Improving weighted least squares inference

DiCiccio, Cyrus J ; Romano, Joseph P ; Wolf, Michael

Abstract: These days, it is common practice to base inference about the coefficients in a hetoskedastic linear model on the ordinary least squares estimator in conjunction with using heteroskedasticity consistent standard errors. Even when the true form of heteroskedasticity is unknown, heteroskedasticity consistent standard errors can also used to base valid inference on a weighted least squares estimator and using such an estimator can provide large gains in efficiency over the ordinary least squares estimator. However, intervals based on asymptotic approximations with plug-in standard errors often have coverage that is below the nominal level, especially for small sample sizes. Similarly, tests can have null rejection probabilities that are above the nominal level. It is shown that under unknown hereroskedasticy, a bootstrap approximation to the sampling distribution of the weighted least squares estimator is valid, which allows for inference with improved finite-sample properties. For testing linear constraints, permutations tests are proposed which are exact when the error distribution is symmetric and is asymptotically valid otherwise. Another concern that has discouraged the use of weighting is that the weighted least squares estimator may be less efficient than the ordinary least squares estimator when the model used to estimate the unknown form of the heteroskedasticity is misspecified. To address this problem, a new estimator is proposed that is asymptotically at least as efficient as both the ordinary and the weighted least squares estimator. Simulation studies demonstrate the attractive finite-sample properties of this new estimator as well as the improvements in performance realized by bootstrap confidence intervals.

DOI: https://doi.org/10.1016/j.ecosta.2018.06.005

University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 232

Improving Weighted Least Squares Inference

Cyrus J. DiCiccio, Joseph P. Romano and Michael Wolf

Revised version, November 2017

Improving Weighted Least Squares Inference

Cyrus J. DiCiccio
Department of Statistics
Stanford University
cyrusd@stanford.edu

Joseph P. Romano
Departments of Statistics and Economics
Stanford University
romano@stanford.edu

Michael Wolf
Department of Economics
University of Zurich
michael.wolf@econ.uzh.ch

November 13, 2017

Abstract

These days, it is common practice to base inference about the coefficients in a hetoskedastic linear model on the ordinary least squares estimator in conjunction with using heteroskedasticity consistent standard errors. Even when the true form of heteroskedasticity is unknown, heteroskedasticity consistent standard errors can also used to base valid inference on a weighted least squares estimator and using such an estimator can provide large gains in efficiency over the ordinary least squares estimator. However, intervals based on asymptotic approximations with plug-in standard errors often have coverage that is below the nominal level, especially for small sample sizes. Similarly, tests can have null rejection probabilities that are above the nominal level. In this paper, it is shown that under unknown hereroskedasticy, a bootstrap approximation to the sampling distribution of the weighted least squares estimator is valid, which allows for inference with improved finite-sample properties. For testing linear constraints, permutations tests are proposed which are exact when the error distribution is symmetric and is asymptotically valid otherwise. Another concern that has discouraged the use of weighting is that the weighted least squares estimator may be less efficient than the ordinary least squares estimator when the model used to estimate the unknown form of the heteroskedasticity is misspecified. To address this problem, a new estimator is proposed that is asymptotically at least as efficient as both the ordinary and the weighted least squares estimator. Simulation studies demonstrate the attractive finite-sample properties of this new estimator as well as the improvements in performance realized by bootstrap confidence intervals.

KEY WORDS: Bootstrap, conditional heteroskedasticity, HC standard errors.

JEL classification codes: C12, C13, C21.

1 Introduction

In this paper, we consider the problem of inference in a linear regression model. Under conditional homoskedasticity, the ordinary least squares (OLS) estimator is the best linear unbiased estimator. Traditional inference based upon the ordinary least squares estimator, such as the F test or t confidence intervals for individual coefficients, relies on estimators of asymptotic variance that are only consistent when the model is conditionally homoskedastic. In many applications, the assumption of conditional homoskedasticity is unrealistic. When instead the model exhibits conditional heteroskedasticity, traditional inference based on the ordinary least squares estimator may fail to be valid, even asymptotically.

If the skedastic function is known (that is, the function that determines the conditional heteroskedasticty of the error term given the values of the regressors), the best linear unbiased estimator (BLUE) is obtained by computing the ordinary least squares estimator after weighting the data by the inverse of square root of the value of the skedastic function. Unfortunately, in all but the most ideal examples, the heteroskedasticity is of unknown form, and this estimator cannot be used. However, if the skedastic function can be estimated, then weighting the model by the inverse square root of the estimate of the skedastic function produces a "feasible" weighted least squares (WLS) estimator. Although this estimator is no longer unbiased, it can often give improvements in efficiency over the ordinary least squares estimator. Even so, estimating the skedastic function is often challenging, and a poorly estimated skedastic function may produce an estimator that is less efficient than the ordinary least squares estimator. Furthermore, when the estimated skedastic function is not consistent, traditional inference based on the weighted least squares estimator may not be valid. Because of these difficulties the weighted least squares estimator has largely fallen out of favor with practitioners.

As an alternative, White (1980) develops heteroskedasticity consistent (HC) standard errors which allow for asymptotically valid inference, based on the ordinary least squares estimator, in the presence of conditional heteroskedasticity of unknown form. Although this approach abandons any efficiency gains that could be achieved from weighting, the standard errors are consistent under minimal model assumptions. Despite the asymptotic validity, simulation studies, such as MacKinnon and White (1985) who investigate the performance of several different heteroskedasticity consistent standard errors, show that inference based on normal or even t approximations can be misleading in small samples. In such cases, it is useful to consider bootstrap methods.

Following the proposal of White's heteroskedasticity consistent covariance estimators, resampling methods have been developed that give valid inference based on the ordinary least squares estimator. Freedman (1981) proposes the pairs bootstrap which resamples pairs of predictor and response variables from the original data. Another popular technique is the wild bootstrap which is suggested by Wu (1986). This method generates bootstrap samples by simulating error terms according to a distribution whose variance is an estimate of the conditional variance for each predictor variable.

The choice of distribution used to simulate the error terms is discussed extensively in Davidson and Flachaire (2008), Chesher (1989), and MacKinnon (2012). Recent numerical work comparing the pairs bootstrap and the wild bootstrap to asymptotic approximations is given in Flachaire (2005) and Cribari-Neto (2004). Godfrey and Orne (2004) conducts simulations suggesting that combining heteroskedasticity consistent standard errors with the wild bootstrap produces tests that are more reliable in small samples than using the normal approximation. Despite the improvements that the resampling methods produce over asymptotic approximations, inference based on the ordinary least squares estimator may still not be as efficient as weighted least squares.

Neither the solution of using heteroscedasticity consistent covariance estimators, nor using weighted least squares with traditional inference seem entirely satisfactory. Even recently there has been debate about the merits of weighting. Angrist and Pischke (2010) are of the belief that any potential efficiency gains from using a weighted least squares estimator are not substantial enough to risk the harm that could be done by poorly estimated weights. On the other hand, Leamer (2010) contends that researchers should be working to model the heteroskedasticity in order to determine whether sensible reweighting changes estimates or confidence intervals.

Even in examples where the estimated skedastic function is not consistent for the true skedastic function, the weighted least squares estimator can be more efficient than the ordinary least squares estimator in a first order asymptotic sense. Arguably, a more satisfying approach to inference than simply abandoning weighting is to base inference on the weighted least squares estimator in conjunction with HC errors. This proposal goes back to at least Wooldridge (2012) and is made rigorous in Romano and Wolf (2017). Regardless of whether or not the parametric family used to estimate the skedastic function is correctly specified, the weighted least squares estimator has an asymptotically normal distribution with mean zero and a variance that can be consistently estimated by the means of HC standard errors (as long as some mild technical conditions are satisfied).

There are two difficulties with basing inference on these consistent standard errors. As is the case with using White's standard errors, using heteroskedasticity consistent standard errors with the weighted least squares estimator leads to inference that can be misleading in small samples. This problem is even more severe with the weighted estimator than with the ordinary least squares estimator because the plug-in standard errors use the estimated skedastic function, and are the same estimators that would be used if it had been known *a priori* that the model would be weighted by this particular estimated skedastic function. Confidence intervals, for example, do not account for the randomness in estimating the skedastic function and for this reason tend to have coverage that is below the nominal level, especially in small samples.

The other trouble is that inference based on the weighted least squares estimator using consistent standard errors may not be particularly efficient, and investing effort in modeling the conditional variance may be counterproductive. In fact, when the family of skedastic functions is misspecified

(or the estimated skedastic function is not consistent for the true skedastic function), the weighted least squares estimator can be less efficient than the ordinary least squares estimator, even when conditional heteroskedasticity is present. Although this possibility seems rare, it is theoretically unsatisfying and has been given as a reason to abandon the approach altogether.

In this paper, we will address these limitations of the weighted least squares estimator, namely the unsatisfying finite sample performance of asymptotic approximations, and the potential asymptotic inefficiency relative to the ordinary least squares estimator. Thus, the general goal is to improve the methodology in Romano and Wolf (2017) by constructing methods with improved accuracy and efficiency. We begin by establishing that the wild and pairs bootstrap approximations to the sampling distribution of the weighted least squares estimator are consistent. Using resampling methods, rather than asymptotic approximations, has the advantage that for each resample, the skedastic function can be re-estimated. This leads to approximations of the sampling distribution which account for the variability from estimating the weights that can have better finite sample properties than asymptotic approximations, which are the same as if the weights had been specified in advance and were non-random. This allows for confidence intervals and hypothesis tests with better finite sample performance than t intervals or F tests. For testing, we further establish asymptotic validity of permutation tests, which also have the advantage of re-estimating the function, but have the added benefit of finite sample exactness in some circumstances. To address the concern of the possible inefficiency of the weighted least squares estimator, we propose a new estimator that is a convex-combination of the ordinary least squares estimator and the weighted least squares estimator and is at least as efficient (asymptotically) as both the weighted and the ordinary least squares estimator (and potentially more efficient than either).

The remainder of the paper is organized as follows. Model assumptions are given in Section 2. Consistency of both the pairs and wild bootstrap approximations to the distribution of the weighted least squares estimator is given in Section 3; notably, the bootstrap accounts for estimation of the skedastic function as it is re-estimated in each bootstrap sample. Tests for linear constraints of the coefficient vector using both bootstrap methods, as well as a randomization test, are given in Section 3.2. Estimators based on a convex-combination of the ordinary and weighted least squares estimators that are asymptotically no worse, but potentially more efficient than the ordinary least squares estimator, as well as the consistency of the bootstrap distribution of these estimators, are given in Section 4. Here, the bootstrap is useful not only to account for the randomness in the skedastic function but also the randomness in the convex weights. Section 5 provides an example where the convex-combination estimator is strictly more efficient than either the ordinary or weighted least squares estimators. Simulations to examine finite-sample performance, as well as an empirical application, are provided in Section 6. Proofs are given in the appendix.

2 Model and Notation

Throughout the paper, we will be concerned with the heteroskedastic linear regression model specified by the following assumptions.

(A1) The model can be written

$$y_i = x_i^\top \beta + \varepsilon_i \ ,$$

$i = 1, ..., n$, where $x_i \in \mathbb{R}^p$ is a vector of predictor variables, and ε_i is an unobservable error term with properties specified below.

(A2) $\{(y_i, x_i)\}$ are independent and identically distributed (i.i.d.) according to a distribution P.

(A3) The error terms have conditional mean zero given the predictor variables:

$$\mathbb{E}(\varepsilon_i | x_i) = 0 \ .$$

(A4) $\Sigma_{xx} \coloneqq \mathbb{E}(x_i x_i^\top)$ is nonsingular.

(A5) $\Omega \coloneqq \mathbb{E}(\varepsilon_i^2 x_i x_i^\top)$ is nonsingular.

(A6) There exists a function $v(\cdot)$, called the skedastic function, such that

$$\mathbb{E}(\varepsilon_i^2 | x_i) = v(x_i) \ .$$

It is also convenient to write the linear model specified by assumption (A1) in vector-matrix notation.

$$Y = X\beta + \varepsilon$$

where

$$Y \coloneqq \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \ , \quad \varepsilon \coloneqq \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \ , \quad \text{and} \quad X \coloneqq \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \dots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix} \ .$$

Finally, following the notation of Romano and Wolf (2017), define

$$\Omega_{a/b} \coloneqq \mathbb{E}\left(x_i x_i^\top \frac{a(x_i)}{b(x_i)} \right)$$

for any functions $a, b : \mathbb{R}^p \to \mathbb{R}$ such that this expectation is finite. Using this convention, $\Sigma_{xx} = \Omega_{1/1}$ and $\Omega = \Omega_{v/1}$.

3 Estimators and Consistency of the Bootstrap

Under the model assumptions given in Section 2, it is common to use the ordinary least squares (OLS) estimator

$$\hat{\beta}_{\text{OLS}} := \left(X^\top X\right)^{-1} X^\top Y$$

to estimate β. Although this estimator is unbiased, it is not efficient when the model is not conditionally homoskedastic. Ideally, one would use the best linear unbiased estimator (BLUE) which is obtained by regressing $y_i/\sqrt{v(x_i)}$ on $x_i/\sqrt{v(x_i)}$ by OLS. But this estimator requires knowledge of the true skedastic function and thus is not feasible in most applications.

Instead, one can estimate the skedastic function and weight the observations by the estimate of the skedastic function. Typically, the skedastic function is estimated by $v_{\hat{\theta}}(\cdot)$, a member of a parametric family $\{v_\theta(\cdot) : \theta \in \mathbb{R}^d\}$ of skedastic functions. For instance, a popular choice for the family of skedastic functions is

$$v_\theta(x_i) := \exp\big(\theta_0 + \theta_1 \log|x_{i,1}| + \ldots + \theta_p \log|x_{i,p}|\big) \ , \quad \text{with} \quad \theta := (\theta_0, \theta_1, \ldots, \theta_p) \in \mathbb{R}^{p+1} \ . \tag{3.1}$$

The weighted least squares (WLS) estimator based on the estimated skedastic function is obtained by regressing $y_i/\sqrt{v_{\hat{\theta}}(x_i)}$ on $x_i/\sqrt{v_{\hat{\theta}}(x_i)}$ by OLS and thus given by

$$\hat{\beta}_{\text{WLS}} := (X^\top V_{\hat{\theta}}^{-1} X)^{-1} X^\top V_{\hat{\theta}}^{-1} Y$$

where $V_\theta := \text{diag}\{v_\theta(x_1), ..., v_\theta(x_n)\}$.

Provided the estimated skedastic function $v_{\hat{\theta}}(\cdot)$ is suitably close to some limiting estimated skedastic function, say $v_{\theta_0}(\cdot)$ for n large, then the weighted least squares estimator has an asymptotically normal distribution. Note that $v_{\theta_0}(\cdot)$ need not correspond to the true skedastic function, which of course happens if the family of skedastic functions is not well specified. Romano and Wolf (2017) assume that $\hat{\theta}$ is a consistent estimator of some θ_0 in the sense that

$$n^{1/4}(\hat{\theta} - \theta_0) \xrightarrow{P} 0 \ , \tag{3.2}$$

where $\xrightarrow{P}$ denotes convergence in probability. This condition is verified by Romano and Wolf (2017) for the family of skedastic functions given in Lemma 3.1 under moment conditions. They also assume that at this θ_0, $1/v_\theta(\cdot)$ is differentiable in the sense that there exists a d-dimensional vector-valued function

$$r_{\theta_0}(x) = \big(r_{\theta_0,1}(x), \ldots, r_{\theta_0,d}(x)\big)$$

and a real-valued function $s_{\theta_0}(\cdot)$ (satisfying some moment assumptions) such that

$$\left|\frac{1}{v_\theta(x)} - \frac{1}{v_{\theta_0}(x)} - r_{\theta_0}(x)(\theta - \theta_0)\right| \leq \frac{1}{2}|\theta - \theta_0|^2 s_{\theta_0}(x) \ , \tag{3.3}$$

for all θ in some small open ball around θ_0 and all x.

If (3.2) and (3.3) are satisfied, then under some further regularity conditions,

$$\sqrt{n}\left(\hat{\beta}_{\mathrm{WLS}} - \beta\right) \xrightarrow{d} N(0, \Omega_{1/w}^{-1}\Omega_{v/w^2}\Omega_{1/w}^{-1})$$

where $w(\cdot) := v_{\theta_0}(\cdot)$ and $\xrightarrow{d}$ denotes convergence in distribution.

The matrices $\Omega_{1/w}$ and Ω_{v/w^2} appearing in the asymptotic variance can be consistently estimated by

$$\hat{\Omega}_{1/w} := \frac{X' V_{\hat{\theta}}^{-1} X}{n}$$

and

$$\hat{\Omega}_{v/w^2} := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\tilde{\varepsilon}_i^2}{v_{\hat{\theta}}^2(x_i)} \cdot x_i x_i^\top\right) ,$$

respectively, for suitable residuals $\tilde{\varepsilon}$ that are consistent for the true error terms ε. Then the asymptotic variance of the weighted least squares estimator, denoted by $\mathrm{Avar}(\hat{\beta}_{\mathrm{WLS}})$, can be consistently estimated by

$$\widehat{\mathrm{Avar}}\left(\hat{\beta}_{\mathrm{WLS}}\right) = \hat{\Omega}_{1/w}^{-1}\hat{\Omega}_{v/w^2}\hat{\Omega}_{1/w}^{-1} . \tag{3.4}$$

Remark 3.1. When the 'raw' OLS residuals, $\hat{\varepsilon}_i := y_i - x_i\hat{\beta}_{\mathrm{OLS}}$, are used to compute $\hat{\Omega}_{v/w^2}$, the estimator (3.4) is commonly referred to as the HC0 estimator. To improve finite-sample performance other variants of HC used scaled residuals instead. The HC1 estimator scales the OLS residuals by $\sqrt{n/(n-p)}$, which reduces bias. When the errors are homoskedastic, the variance of the OLS residual $\hat{\varepsilon}_i$ is proportional to $1/(1-h_i)$, where h_i is the i^{th} diagonal entry of the 'hat' matrix $H := X(X^\top X)^{-1}X^\top$. The HC2 estimator uses the OLS residuals scaled by $1/\sqrt{(1-h_i)}$. The HC3 estimator uses the OLS residuals scaled by $1/(1-h_i)$. ■

3.1 Confidence Intervals

Using the plug-in estimator of asymptotic variance, $\widehat{\mathrm{Avar}}\left(\hat{\beta}_{\mathrm{WLS}}\right)$ in (3.4), gives approximate t confidence intervals for the coefficients having the form

$$\hat{\beta}_{\mathrm{WLS},k} \pm t_{n-p,1-\alpha/2} \cdot \mathrm{SE}(\hat{\beta}_{\mathrm{WLS},k})$$

where

$$\mathrm{SE}(\hat{\beta}_{\mathrm{WLS},k}) := \sqrt{\widehat{\mathrm{Avar}}(\hat{\beta}_{\mathrm{WLS},k})/n} ,$$

and $t_{n-p,1-\alpha/2}$ is the $1-\alpha/2$ quantile of the t-distribution with $n-p$ degrees of freedom. These intervals are asymptotically valid; however, simulations suggest that the true coverage rates are often smaller than the nominal level, especially in small samples. The standard errors for these confidence intervals are the same standard errors that would be used if we had known before observing any data that the model would be weighted by $1/\sqrt{v_{\hat{\theta}}(\cdot)}$ and the intervals do not account for variability in the estimation of the skedastic function. The coverage can be improved by reporting intervals

based on the "pairs" bootstrap confidence intervals where the skedastic function is estimated on each bootstrap sample separately.

The empirical distribution of a sample $(x_1, y_1), ..., (x_n, y_n)$ is

$$\hat{P}_n(s_1, ..., s_p, t) := \frac{1}{n}\sum_{i=1}^{n} I\left\{x_{i,1} \leq s_1, ..., x_{i,p} \leq s_p, y_i \leq t\right\} \ .$$

The pairs bootstrap, which is commonly used for heteroskedastic regression models, generates bootstrap samples, $(x_1^*, y_1^*), ..., (x_n^*, y_n^*)$ from $\hat{P}_n$. Alternatively, one could generate bootstrap samples $(x_1, y_1^*), ..., (x_n, y_n^*)$ using the wild bootstrap which simulates new response variables

$$y_i^* := x_i\hat{\beta}_{\text{WLS}} + \varepsilon_i^*$$

where ε_i^* are sampled from any distribution F with mean zero and variance $\hat{\varepsilon}_i^2$.

Remark 3.2. Typically $\varepsilon_i^* := u_i \cdot \hat{\varepsilon}_i$ where u_i is a random variable with mean zero and variance one. When the errors are symmetric, a commonly used distribution (which will be referred to as the F_2 distribution) for u_i takes values ± 1, each with probability $1/2$. For skewed errors, Mammen (1993) proposes simulating u_i according to a distribution (which will be referred to as the F_1 distribution) that takes values $-(\sqrt{5}-1)/2$ with probability $(\sqrt{5}+1)/(2\sqrt{5})$ and $(\sqrt{5}+1)/2$ with probability $(\sqrt{5}-1)/(2\sqrt{5})$. This distribution has third moment one, and accounts for skewness in the distribution of the errors. ■

When computing the weighted least squares estimator $\hat{\beta}_{\text{WLS}}$, the parameter for the estimated skedastic function is re-estimated on the bootstrap sample by $\hat{\theta}^*$. The following theorem establishes that the distribution of $\sqrt{n}\big(\hat{\beta}^*_{\text{WLS}} - \hat{\beta}_{\text{WLS}}\big)$, using the pairs or the wild bootstrap, is a consistent approximation of the sampling distribution of $\sqrt{n}\big(\hat{\beta}_{\text{WLS}} - \beta\big)$.

Theorem 3.1. *Suppose that $(x_1, y_1), ..., (x_n, y_n)$ are i.i.d. satisfying assumptions $(A1)$–$(A6)$ above, and that $\left\{v_\theta(\cdot) : \theta \in \mathbb{R}^d\right\}$ is a family of continuous skedastic functions satisfying (3.3) for some θ_0 for any functions $r_{\theta_0}(\cdot)$ and $s_{\theta_0}(\cdot)$ such that*

$$\mathbb{E}\left|x_1 y_1 r(x_1)\right|^2 < \infty \quad \textit{and} \quad \mathbb{E}\left|x_1 y_1 s(x_1)\right|^2 < \infty \ .$$

Let $\hat{\theta}$ be an estimator satisfying (3.2). Further suppose that $n^{1/4}\big(\hat{\theta}^* - \theta_0\big)$ converges to zero in conditional probability. (These assumptions are verified, under moment assumptions, for a particular parametric family of skedastic functions in Lemma 3.1). Let $\hat{\beta}_{WLS} := (X^\top V_{\hat{\theta}}^{-1} X)^{-1} X^\top V_{\hat{\theta}}^{-1} Y$ and $v_{\theta_0} =: w$ so that $W = Diag(v_{\theta_0}(x_1), ..., v_{\theta_0}(x_n))$. If

$$\mathbb{E}\left(\frac{\left(y_i^2 + \sum_{j=1}^{p} x_{i,j}^2\right)^2}{w^2(x_i)}\right) < \infty \ ,$$

Ω_{v/w^2} and $\Omega_{1/w}$ exist, and $\Omega_{1/w}$ is invertible, then the conditional law of $\sqrt{n}\big(\hat{\beta}^*_{WLS} - \hat{\beta}_{WLS}\big)$, based on a pairs bootstrap sample or a wild bootstrap sample, converges weakly to the multivariate normal distribution with mean zero and covariance matrix $\Omega^{-1}_{1/w}\Omega_{v/w^2}\Omega^{-1}_{1/w}$ in probability. Furthermore, for any k, the distribution of $\sqrt{n}\big(\hat{\beta}^*_{WLS,k} - \hat{\beta}_{WLS,k}\big)/\sqrt{\widehat{Avar}(\hat{\beta}_{WLS,k})^*}$ is asymptotically standard normal in probability, where $\sqrt{\widehat{Avar}(\hat{\beta}_{WLS,k})^*/n}$ is the estimated standard error of $\hat{\beta}^*_{WLS,k}$ using the bootstrap sample.

Remark 3.3. Of course, the bootstrap distribution is random and hence its weak convergence properties hold in a probabilistic sense. As is customary, when we say that a sequence of random distributions, say $\hat{G}_n$ converges weakly to G in probability, we mean that $\rho(\hat{G}_n, G) \xrightarrow{P} 0$ where ρ is any metric metrizing weak convergence on the space of distributions. We also say that a sequence $T_n(X, Y)$ converges in conditional probability to zero almost surely if for almost every sequence $\{x_i, y_i\}$, $T_n(X^*, Y^*) \to 0$ in $\hat{P}_n$ probability. ■

The approximation given in Theorem 3.1 guarantees the basic bootstrap confidence intervals computed by

$$\Big(\hat{\beta}_{\text{WLS},k} - q(1-\alpha/2, \hat{P}), \hat{\beta}_{\text{WLS},k} - q(\alpha/2, \hat{P})\Big)$$

are asymptotically level α, where $q(\alpha, \hat{P})$ denotes the α quantile of

$$\sqrt{n}\big(\hat{\beta}^*_{\text{WLS},k} - \hat{\beta}_{\text{WLS},k}\big) \ .$$

Rather than using the basic bootstrap confidence intervals, bootstrap-t intervals can be constructed. Again appealing to 3.1, the bootstrap-t intervals

$$\left(\hat{\beta}_{\text{WLS}} - \sqrt{\widehat{\text{Avar}}(\hat{\beta}_{\text{WLS},k})/n} \cdot t(1-\alpha/2, \hat{P}), \hat{\beta}_{\text{WLS}} - \sqrt{\widehat{\text{Avar}}(\hat{\beta}_{\text{WLS},k})/n} \cdot t(\alpha/2, \hat{P})\right)$$

are asymptotically level α where $t(\alpha, \hat{P})$ denotes the α quantile of

$$\frac{\sqrt{n}\big(\hat{\beta}^*_{\text{WLS},k} - \hat{\beta}_{\text{WLS},k}\big)}{\sqrt{\widehat{\text{Avar}}(\hat{\beta}_{\text{WLS},k})^*}} \ .$$

Remark 3.4 (Adaptive Least Squares)**.** Romano and Wolf (2017) propose choosing between the OLS and WLS estimators by applying a test for conditional heteroskedasticity and call the resulting estimator the adaptive least squares (ALS) estimator. The confidence intervals reported for the ALS estimator, agree with either the confidence intervals for the WLS or OLS estimators (using HC standard errors), depending on the decision of the test. Rather than using asymptotic intervals, the corresponding bootstrap intervals for either the WLS or OLS estimators can be used for the ALS estimator. ■

In Theorem 3.1, it was assumed that we have a family of skedastic functions $\{v_\theta(\cdot)\}$, and an estimator of θ, say $\hat{\theta}$, such that $n^{1/4}\big(\hat{\theta}^* - \theta_0\big)$ converges in conditional probability to zero. We will

now verify this assumption for a flexible family of skedastic functions which includes the family specified in (3.1).

Lemma 3.1. *For any functions $g_i : \mathbb{R}^d \to \mathbb{R}^d$, $i = 1, ..., d$, define the family $\left\{v_\theta : \theta \in \mathbb{R}^d\right\}$ by*

$$v_\theta(x) := \exp\left[\sum_{i=1}^{d} \theta_j g_j(x)\right] ,$$

and let $\hat{\theta}$ be the estimator obtained by regressing $h_\delta(\hat{\varepsilon}_i) := \log\left(\max\left\{\delta^2, \hat{\varepsilon}_i^2\right\}\right)$ (where $\hat{\varepsilon}_i := y_i - x_i\hat{\beta}_{OLS}$) on $g(x_i) = \left(g_1(x_i), ..., g_d(x_i)\right)$ by OLS, where $\delta > 0$ is a small constant. Then, $n^{1/4}\left(\hat{\theta}^* - \theta_0\right)$ converges in conditional probability to zero for

$$\theta_0 := E(g(x_i)g(x_i)')E(g(x_i)h_\delta(\varepsilon_i))$$

provided $E(g_j(x_i)g_k(x_i))^{4/3}$ and $E(g_j(x_i)h_\delta(\varepsilon_i))^{4/3}$ are both finite for each j and k.

3.2 Hypothesis Testing

Just as using a t approximation often produces confidence intervals with coverage below the nominal confidence level, especially for small samples, using an F approximation to conduct F tests of linear constraints often gives rejection probabilities that are above the nominal significance level, especially for small samples. And as with confidence intervals, using the bootstrap can produce tests that have rejection probabilities that are closer to the nominal level. Consider the hypothesis

$$H_0 : R\beta = q$$

where R is a $J \times p$ matrix of full rank (with $J \leq p$) and q is a vector of length J. Two appropriate test statistics for this hypothesis are the Wald statistic

$$W_n(X, Y) := n \cdot \left(R\hat{\beta}_{\text{WLS}} - q\right)^\top \left[R\hat{\Omega}_{1/w}^{-1}\hat{\Omega}_{v/w^2}\hat{\Omega}_{1/w}^{-1}R^\top\right]^{-1} \left(R\hat{\beta}_{\text{WLS}} - q\right) , \tag{3.5}$$

and the maximum statistic,

$$M_n(X, Y) := \max_{1 \leq k \leq p} \left\{ \frac{\left|[R\hat{\beta}_{\text{WLS}}]_k - q_k\right|}{\left[R\widehat{\Omega}_{1/w}^{-1}\hat{\Omega}_{v/w^2}\hat{\Omega}_{1/w}^{-1}R^\top\right]_{k,k}} \right\} . \tag{3.6}$$

It follows immediately from the results of Romano and Wolf (2017) that, under the null, the sampling distribution of $W_n(X, Y)$ is asymptotically chi-squared with J degrees of freedom and the sampling distribution of $M_n(X, Y)$ is asymptotically distributed as the maximum of the absolute values of k correlated standard normal variables. Let $G_n(x, P)$ denote the sampling distribution of W_n when (X_1, Y_1) are distributed according to P.

Define $c_n(1 - \alpha, \hat{P})$ to be the $1 - \alpha$ quantile of the distribution of

$$\left(R\left(\hat{\beta}^*_{\text{WLS}} - \hat{\beta}_{\text{WLS}}\right)\right)^\top \left[R\hat{\Omega}^{*-1}_{1/w}\hat{\Omega}^*_{v/w^2}\hat{\Omega}^{*-1}_{1/w}R^\top\right]^{-1} \left(R\left(\hat{\beta}^*_{\text{WLS}} - \hat{\beta}_{\text{WLS}}\right)\right)$$

and $d_n(1-\alpha, \hat{P})$ to be the $1-\alpha$ quantile of the distribution of

$$\max_{1\leq k\leq p}\left\{\frac{\left([R\hat{\beta}^*_{\text{WLS}}]_k - [R\hat{\beta}_{\text{WLS}}]_k\right)}{\left[R\hat{\Omega}^{*-1}_{1/w}\hat{\Omega}^*_{v/w^2}\hat{\Omega}^{*-1}_{1/w}R^\top\right]_{k,k}}\right\}$$

using the pairs or wild bootstrap.

Theorem 3.2. *Suppose that $(x_1, y_1), ..., (x_n, y_n)$ are i.i.d. according to a distribution P such that $R\beta = q$. Then, under the assumptions of Theorem 3.1,*

$$P\left(W_n(X, Y) > c_n(1-\alpha, \hat{P}_n)\right) \to \alpha$$

as $n \to \infty$. That is, the bootstrap quantiles of the Wald statistic converge to the corresponding quantiles of a chi-squared distribution with J degrees of freedom when $R\beta = q$. Similarly,

$$P\left(M_n(X, Y) > d_n(1-\alpha, \hat{P}_n)\right) \to \alpha$$

as $n \to \infty$.

We point out that hypothesis testing using the wild bootstrap is closely related to a commonly used randomization test under symmetry assumptions.

Suppose that the ε_i follow a symmetric distribution conditional on X_i in the sense that the distribution of ε_i given X_i is the same as the distribution of $-\varepsilon_i$ given X_i. Then under $H : \beta = 0$, the joint distribution of the (X_i, Y_i) is invariant under the group of transformations $\mathbf{G}_n := \{g_\delta : \delta \in \{1, -1\}^n\}$ such that $g_\delta((x_1, y_1), ..., (x_n, y_n)) = ((x_1, \delta_1 y_1), ..., (x_n, \delta_n y_n))$ for any $x, y \in \mathbb{R}^n$. Given a test statistic T_n used to test the hypothesis $H : \beta = 0$, the permutation test rejects if $T_n(X, Y)$ exceeds the appropriate quantiles of the permutation distribution of T_n, which is given by

$$\hat{R}^{T_n}_n(t) := \frac{1}{2^n}\sum_{g_\delta \in \mathbf{G}_n} I\left\{T_n(X, g_\delta(Y)) \leq t\right\} .$$

For any choice of test statistic, the invariance of the distribution of the data under the group of transformations is sufficient to ensure that the randomization test is exact; see Lehmann and Romano (2005, Chapter 15) for details.

Typically for regression problems, the test statistic is chosen to be the usual F statistic in homoskedastic models, or the Wald statistic in heteroskedastic models. While under the symmetry assumption this test is exact in either setting, Janssen (1999) shows that this test is robust against violations of the symmetry assumptions (in the sense that the test is still asymptotically valid when the distribution of the Y_i is not symmetric).

When the symmetry assumption is satisfied, the randomization test using W_n or M_n — as defined in equations (3.5) and (3.6), respectively — is exact in the sense that the null rejection probability is exactly the nominal level for any sample size. Even when this assumption is not satisfied, the test is still asymptotically valid, as the following theorem demonstrates.

Theorem 3.3. *Suppose that $(x_1, y_1), ..., (x_n, y_n)$ are i.i.d. according to a distribution P such that $\beta = 0$. Suppose that $n^{1/4}(\hat{\theta}(g_\delta(X, Y)) - \theta_0)$ converges in probability to zero conditionally on the X's and Y's for any uniformly randomly chosen $g_\delta \in \mathbb{G}_n$. (This assumption is verified, under moment assumptions, for a particular parametric family of skedastic functions in Lemma 3.2). Then, under the assumptions of Theorem 3.1, the permutation distribution $\hat{R}_n^{W_n}$ of W_n satisfies*

$$\sup_{t \in \mathbb{R}} \left| \hat{R}_n^{W_n}(t) - J_n^{W_n}(t, P) \right| \to 0$$

in probability as $n \to \infty$ where $J_n^{W_n}(\cdot, P)$ is the sampling distribution of W_n under P. Similarly, the permutation distribution $\hat{R}_n^{M_n}$ of M_n satisfies

$$\sup_{t \in \mathbb{R}} \left| \hat{R}_n^{M_n}(t) - J_n^{M_n}(t, P) \right| \to 0$$

in probability as $n \to \infty$ where $J_n^{M_n}(\cdot, P)$ is the sampling distribution of M_n under P.

Once again, this theorem makes assumptions about the convergence in probability of the estimate of the parameter in the skedastic function. We verify this assumption for a particular family of skedastic functions.

Lemma 3.2. *For any functions $g_i : \mathbb{R}^d \to \mathbb{R}^d$, $i = 1, ..., d$, define the family $\left\{v_\theta : \theta \in \mathbb{R}^d\right\}$ by*

$$v_\theta(x) := \exp\left[\sum_{i=1}^{d} \theta_j g_j(x)\right] ,$$

and let $\hat{\theta}$ be the estimator obtained by regressing $h_\delta(\hat{\varepsilon}_i) := \log\left(\max\left\{\delta^2, \hat{\varepsilon}_i^2\right\}\right)$ on $g(x_i) = \left(g_1(x_i), ..., g_d(x_i)\right)$ by OLS, where $\delta > 0$ is a small constant. Then, for any randomly and uniformly chosen $g_\delta \in \mathbb{G}_n$, $n^{1/4}\left(\hat{\theta}(g_\delta(X, Y)) - \theta_0\right)$ converges in conditional probability to zero for

$$\theta_0 := E(g(x_i)g(x_i)')E(g(x_i)h_\delta(\varepsilon_i))$$

provided $E(g_j(x_i)g_k(x_i))^{4/3}$ and $E(g_j(x_i)h_\delta(\varepsilon_i))^{4/3}$ are both finite for each j and k.

4 A Convex Linear Combination of the Ordinary and Weighted Least Squares Estimators

When the family of skedastic functions is misspecified, the weighted least squares estimator can be less efficient than the ordinary least squares estimator, even asymptotically.

When interested in inference for a particular coefficient, say β_k, practitioners might be tempted to decide between the ordinary and weighted least squares estimators based on which estimator has the smaller standard error In particular, it might be tempting to report the estimator

$$\hat{\beta}_{\text{MIN},k} := \begin{cases} \hat{\beta}_{\text{WLS},k} & \text{if} \quad \widehat{\text{Avar}}(\hat{\beta}_{\text{OLS},k}) > \widehat{\text{Avar}}(\hat{\beta}_{\text{WLS},k}) \\ \hat{\beta}_{\text{OLS},k} & \text{if} \quad \widehat{\text{Avar}}(\hat{\beta}_{\text{OLS},k}) \leq \widehat{\text{Avar}}(\hat{\beta}_{\text{WLS},k}) \end{cases} ,$$

along with the corresponding confidence interval

$$\hat{\beta}_{\text{MIN},k} \pm t_{n-p,1-\alpha/2} \cdot \sqrt{\frac{1}{n} \min\left\{\widehat{\text{Avar}}(\hat{\beta}_{\text{WLS},k}), \widehat{\text{Avar}}(\hat{\beta}_{\text{OLS},k})\right\}} \ . \tag{4.1}$$

Asymptotically, this estimator has the same efficiency as the better of the ordinary least squares and weighted estimators. However, the confidence interval (4.1) tends to undercover in finite samples due to the minimizing over the standard error. The next theorem establishes consistency of the bootstrap (and also bootstrap-t) distribution, which can be used to produce confidence intervals with better finite-sample coverage than those given by (4.1).

Theorem 4.1. *Under the conditions of Theorem 3.1, the sampling distribution of* $\sqrt{n}(\hat{\beta}_{MIN,k} - \beta_k)$ *converges weakly to the normal distribution with mean zero and variance*

$$\sigma^2_{MIN} \coloneqq \min\left\{Avar(\hat{\beta}_{WLS,k}), Avar(\hat{\beta}_{OLS,k})\right\}$$

The distribution of $\sqrt{n}(\hat{\beta}^*_{MIN,k} - \hat{\beta}_k)$*, where the samples* (x^*_i, y^*_i) *are generated according to the pairs bootstrap or the wild bootstrap, converges weakly to the normal distribution having mean zero and variance* σ^2_{MIN} *in probability. Furthermore, for any* k*, the distribution of* $\sqrt{n}(\hat{\beta}^*_{MIN,k} - \hat{\beta}_{MIN,k})/\hat{\sigma}^*_{MIN}$ *is asymptotically standard normal in probability, where*

$$\sigma^*_{MIN} \coloneqq \min\left\{\sqrt{Avar(\hat{\beta}_{WLS,k})^*}, \sqrt{Avar(\hat{\beta}_{OLS,k})^*}\right\} \ .$$

When the estimated skedastic function is consistent for the true skedastic function, the estimator $\hat{\beta}_{\text{MIN},k}$ is asymptotically as efficient as the best linear unbiased estimator. On the other hand, when the skedastic function is misspecified, one can find an estimator which is at least as efficient as $\hat{\beta}_{\text{MIN}}$, regardless of whether or not the skedastic function is well modeled, but can potentially have smaller asymptotic variance. With the aim of creating such an estimator, consider estimators of the form

$$\hat{\beta}_\lambda \coloneqq \lambda\hat{\beta}_{\text{OLS}} + (1-\lambda)\hat{\beta}_{\text{WLS}} \tag{4.2}$$

for $\lambda \in [0,1]$, which are convex-combinations of the ordinary and weighted least squares estimators. To study the asymptotic behavior of these estimators, it is helpful to first find the asymptotic joint distribution of the ordinary and weighted least squares estimators.

Theorem 4.2. *Under the assumptions of Theorem 3.1,*

$$\sqrt{n}\left(\begin{pmatrix} \hat{\beta}_{WLS} \\ \hat{\beta}_{OLS} \end{pmatrix} - \begin{pmatrix} \beta \\ \beta \end{pmatrix}\right) \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Omega^{-1}_{1/w}\Omega_{v/w^2}\Omega^{-1}_{1/w} & \Omega^{-1}_{1/w}\Omega_{v/w}\Omega^{-1}_{1/1} \\ \Omega^{-1}_{1/1}\Omega_{v/w}\Omega^{-1}_{1/w} & \Omega^{-1}_{1/1}\Omega_{v/1}\Omega^{-1}_{1/1} \end{pmatrix}\right)$$

as $n \to \infty$ *.*

It follows that for any $\lambda \in [0,1]$, $\sqrt{n}(\hat{\beta}_\lambda - \beta)$ asymptotically has a normal distribution with mean zero and covariance matrix

$$\text{Avar}(\hat{\beta}_\lambda) \coloneqq \lambda^2\Omega^{-1}_{1/w}\Omega_{v/w^2}\Omega^{-1}_{1/w} + 2\lambda(1-\lambda)\Omega^{-1}_{1/w}\Omega_{v/w}\Omega^{-1}_{1/1} + (1-\lambda)^2\Omega^{-1}_{1/1}\Omega_{v/1}\Omega^{-1}_{1/1} \ ,$$

which can be consistently estimated by

$$\widehat{\text{Avar}}(\hat{\beta}_\lambda) := \left[\lambda^2 \hat{\Omega}_{1/w}^{-1} \hat{\Omega}_{v/w^2} \hat{\Omega}_{1/w}^{-1} + 2\lambda(1-\lambda)\hat{\Omega}_{1/w}^{-1} \hat{\Omega}_{v/w} \hat{\Omega}_{1/1}^{-1} + (1-\lambda)^2 \hat{\Omega}_{1/1}^{-1} \hat{\Omega}_{v/1} \hat{\Omega}_{1/1}^{-1}\right] .$$

For any particular coefficient β_k, it then holds that $\sqrt{n}\big(\hat{\beta}_{\lambda,k} - \beta_k\big)$ is asymptotically normal with mean zero and variance $\text{Avar}(\hat{\beta}_{\lambda,k})$, which denotes the k^{th} diagonal entry of $\text{Avar}(\hat{\beta}_\lambda)$. This variance can be consistently estimated by $\widehat{\text{Avar}}(\hat{\beta}_{\lambda,k})$, the k^{th} diagonal entry of $\widehat{\text{Avar}}(\hat{\beta}_\lambda)$. In conjunction with this standard error, the estimator $\hat{\beta}_{\lambda,k}$ can be used for inference about β_k. For instance, asymptotically valid t confidence intervals are given by

$$\hat{\beta}_{\lambda,k} \pm t_{n-p,1-\alpha/2} \cdot \sqrt{\widehat{\text{Avar}}(\hat{\beta}_{\lambda,k})/n} .$$

These intervals suffer from the same shortcomings as the asymptotic confidence intervals based on the weighted least squares estimator. But using the bootstrap can once again lead to improved finite-sample performance, and the following theorem establishes consistency of the bootstrap (and also bootstrap-t) distribution.

Theorem 4.3. *Under the conditions of Theorem 3.1, $\sqrt{n}\big(\hat{\beta}^*_\lambda - \hat{\beta}_\lambda\big)$, using the pairs or the wild bootstrap, converges weakly to the normal distribution with mean zero and variance $Avar(\hat{\beta}_\lambda)$, in probability for any fixed λ. Furthermore, for any k, the distribution of $\sqrt{n}\big(\hat{\beta}^*_{\lambda,k} - \hat{\beta}_{\lambda,k}\big)/\sqrt{\widehat{Avar}(\hat{\beta}_{\lambda,k})^*}$ is asymptotically standard normal in probability, where $\sqrt{\widehat{Avar}(\hat{\beta}_{\lambda,k})^*/n}$ is the estimated standard error of $\hat{\beta}^*_{\lambda,k}$ using the bootstrap sample.*

Although inference for β_k can be based on $\hat{\beta}_\lambda$ for any $\lambda \in [0,1]$, we would like to choose a value of λ that results in an efficient estimator. The asymptotic variance $\text{Avar}(\hat{\beta}_{\lambda,k})$ is a quadratic function of λ, and therefore has a unique minimum, say λ_0, over the interval $[0,1]$ unless $\text{Avar}(\hat{\beta}_{\lambda,k})$ is constant in λ (which may occur if there is homoskedasticity); in this case, define $\lambda_0 = 1$. Asymptotically, $\hat{\beta}_{\lambda_0,k}$ is the most efficient estimate of β_k amongst the collection $\big\{\hat{\beta}_{\lambda,k} : \lambda \in [0,1]\big\}$. Because this collection includes both the weighted and ordinary least squares estimators, $\hat{\beta}_{\lambda_0,k}$ is at least as efficient as the ordinary least squares estimator, and may have considerably smaller asymptotic variance when the skedastic function is well modeled. In fact, this estimator can have smaller asymptotic variance than both the ordinary and weighted least squares estimators. Unfortunately, without knowing the asymptotic variance, we cannot find λ_0 and we cannot use the estimator $\hat{\beta}_{\lambda_0,k}$. Instead, we can estimate λ_0 by $\hat{\lambda}_0$, the minimum of $\widehat{\text{Avar}}(\hat{\beta}_{\lambda,k})$ over the interval $[0,1]$, provided there is a unique minimum (otherwise set $\hat{\lambda}_0 = 1$). In particular, the minimizer is given by

$$\hat{\lambda}_0 = \frac{\left[\hat{\Omega}_{1/1}^{-1} \hat{\Omega}_{v/1} \hat{\Omega}_{1/1}^{-1} - \hat{\Omega}_{1/w}^{-1} \hat{\Omega}_{v/w} \hat{\Omega}_{1/1}^{-1}\right]_{k,k}}{\left[\hat{\Omega}_{1/w}^{-1} \hat{\Omega}_{v/w^2} \hat{\Omega}_{1/w}^{-1} - 2 \cdot \hat{\Omega}_{1/w}^{-1} \hat{\Omega}_{v/w} \hat{\Omega}_{1/1}^{-1} + \hat{\Omega}_{1/1}^{-1} \hat{\Omega}_{v/1} \hat{\Omega}_{1/1}^{-1}\right]_{k,k}} ,$$

if this quantity lies in the interval [0,1], or otherwise $\hat{\lambda}_0$ is zero or one depending on which gives a

smaller variance. If we choose to use the estimator, $\hat{\beta}_{\hat{\lambda}_0,k}$, then the confidence interval

$$\hat{\beta}_{\hat{\lambda}_0,k} \pm t_{n-p,1-\alpha/2} \cdot \sqrt{\frac{1}{n}\widehat{\mathrm{Avar}}(\hat{\beta}_{\hat{\lambda}_0,k})}$$

will tend to have a coverage rate that is (much) smaller than the nominal level in finite samples, since the smallest estimated variance is likely downward biased for the true variance. Instead, reporting bootstrapped confidence intervals where the $\hat{\lambda}_0$ is recomputed for each bootstrap sample may give more reliable confidence intervals. The next theorem demonstrates that the bootstrap distribution of $\sqrt{n}\big(\hat{\beta}^*_{\hat{\lambda}^*_0,k} - \hat{\beta}_{\hat{\lambda}_0,k}\big)$ consistently approximates the sampling distribution of $\sqrt{n}\big(\hat{\beta}_{\hat{\lambda}_0,k} - \beta_k\big)$.

Theorem 4.4. *Under the conditions of Theorem 3.1, the sampling distribution of $\sqrt{n}\big(\hat{\beta}_{\hat{\lambda}_0,k} - \beta_k\big)$ converges weakly to the normal distribution with mean zero and variance $Avar(\hat{\beta}_{\lambda_0,k})$ and the bootstrap distribution of $\sqrt{n}\big(\hat{\beta}^*_{\hat{\lambda}^*_0,k} - \hat{\beta}_{\hat{\lambda}_0,k}\big)$ also converges weakly to the normal distribution with mean zero and variance $Avar(\hat{\beta}_{\lambda_0,k})$ in probability. Also, for any k, the distribution of $\sqrt{n}\big(\hat{\beta}^*_{\hat{\lambda}_0,k} - \hat{\beta}_{\lambda_0,k}\big)/\sqrt{\widehat{Avar}(\hat{\beta}_{\hat{\lambda},k})^*}$ converges to the standard normal distribution in probability, where $\sqrt{\widehat{Avar}(\hat{\beta}_{\hat{\lambda},k})^*/n}$ is the estimated standard error of $\hat{\beta}^*_{\hat{\lambda},k}$ using the bootstrap sample.*

5 Toy Examples of Linear Combinations with Lower Variance

We will now give an example of a regression model where the optimal λ is in $[0, 1]$ followed by an example where the optimal λ is outside of $[0, 1]$.

For both examples, we will consider the simplest case, namely univariate regression through the origin:

$$y_i = \beta x_i + \varepsilon_i \ .$$

For the first example, let x_i be uniform on the interval $[-1, 1]$ and ε_i have conditional mean zero and conditional variance $\mathrm{var}(\varepsilon_i|x_i) = \sqrt{|x_i|}$. In this example, we will estimate the skedastic function from the family $\{v_\theta(x) = \theta \cdot |x| : \theta > 0\}$ by regressing the squared residuals, $\hat{\epsilon}_i^2$ on the $|x_i|$. Consequently,

$$\theta_0 = \mathbb{E}\big(|x_i|^2\big)^{-1}\mathbb{E}\big(|x_i|\varepsilon_i^2\big) = \mathbb{E}\big(|x_i|^2\big)^{-1}\mathbb{E}\big(|x_i|^{3/2}\big) = \frac{6}{5}$$

The estimator $(1-\lambda)\hat{\beta}_{\mathrm{WLS}} + \lambda\hat{\beta}_{\mathrm{OLS}}$ has variance

$$(1-\lambda)^2\frac{\mathbb{E}\sqrt{|x_i|}}{\big(\mathbb{E}\,|x_i|\big)^2} + 2\lambda(1-\lambda)\frac{\mathbb{E}\,|x_i|^{3/2}}{\mathbb{E}\,|x_i|\,\mathbb{E}x_i^2} + \lambda^2\frac{\mathbb{E}\,|x_i|^{5/2}}{\big(\mathbb{E}x_i^2\big)^2} \ ,$$

λ:		0	.25	.50	.75	1	14/23
n = 20	eMSE	0.1449	0.1380	0.1345	0.1344	0.1378	0.1340
	Coverage	0.9613	0.9596	0.9575	0.9553	0.9527	0.9573
	Width	1.6645	1.6267	1.6066	1.6057	1.6247	1.6038
n = 50	eMSE	0.0564	0.0539	0.0527	0.0528	0.0540	0.0525
	Coverage	0.9524	0.9487	0.9465	0.9449	0.9448	0.9465
	Width	0.9589	0.9371	0.9258	0.9253	0.9360	0.9242
n = 100	eMSE	0.0270	0.0259	0.0254	0.0254	0.0261	0.0255
	Coverage	0.9520	0.9514	0.9506	0.9486	0.9481	0.9483
	Width	0.6592	0.6448	0.6375	0.6376	0.6450	0.6366

Table 5.1: Empirical mean squared error of estimators of β as well as coverage and average length of confidence intervals based on the normal approximation.

which is minimized by

$$\begin{aligned}\lambda_0 &= 1 - \frac{-\frac{\mathbb{E}|x_i|^{3/2}}{\mathbb{E}|x_i|\mathbb{E}x_i^2} + \frac{\mathbb{E}|x_i|^{5/2}}{\left(\mathbb{E}x_i^2\right)^2}}{\frac{\mathbb{E}\sqrt{|x_i|}}{\left(\mathbb{E}|x_i|\right)^2} - 2\frac{\mathbb{E}|x_i|^{3/2}}{\mathbb{E}|x_i|\mathbb{E}x_i^2} + \frac{\mathbb{E}|x_i|^{5/2}}{\left(\mathbb{E}x_i^2\right)^2}} \\ &= 1 - \frac{-\frac{12}{5} + \frac{18}{7}}{\frac{8}{3} - 2\frac{12}{5} + \frac{18}{7}} \\ &= \frac{14}{23} \ .\end{aligned}$$

Table 5.1 presents the empirical mean squared error (eMSE) of this estimator for various λ, as well as the coverage and average length of t intervals (with nominal coverage probability 95%) based on 10,000 simulations. For these simulations, the error terms are normally distributed.

For the second example, let the x_i be standard normal, and ε_i have conditional mean zero and conditional variance $\mathrm{var}(\varepsilon_i|x_i) = x_i^2$. For the weighted least squares estimator, we will again use the incorrectly specified family of skedastic functions $\{v_\theta(x) = \theta \cdot |x| : \theta > 0\}$.

In this example, the value of λ minimizing the asymptotic variance of $(1-\lambda)\hat{\beta}_{\mathrm{WLS}} + \lambda\hat{\beta}_{\mathrm{OLS}}$ is

$$\begin{aligned}\lambda_0 &= 1 - \frac{\mathbb{E}\left(x_i^2\right)^{-1}\mathbb{E}\left(x_i^4\right)\mathbb{E}\left(x_i^2\right)^{-1} - \mathbb{E}\left(|x_i|\right)^{-1}\mathbb{E}\left(|x_i|^3\right)\mathbb{E}\left(x_i^2\right)^{-1}}{\mathbb{E}\left(x_i^2\right)^{-1}\mathbb{E}\left(x_i^4\right)\mathbb{E}\left(x_i^2\right)^{-1} - 2 + \mathbb{E}\left(|x_i|\right)^{-1}\mathbb{E}x_i^2\mathbb{E}\left(|x_i|\right)^{-1}} \\ &= 1 - \frac{3-2}{\pi/2 - 4 + 3} \\ &\approx -0.75 \ .\end{aligned}$$

Although choosing values of lambda outside the interval $[0, 1]$ may give estimators with lower variance, we recommend restricting lambda to the interval $[0, 1]$. In situations where $\mathrm{Avar}(\hat{\beta}_\lambda)$

is nearly constant in lambda (such as homoskedastic models), the estimates of λ can be highly unstable when not restricted, and the resulting intervals can have poor coverage. We recommend choosing $\hat{\lambda} = 0$ if the minimizing λ is negative, or $\hat{\lambda} = 1$ if the minimizing λ is positive. Even if the optimal lambda is outside the interval [0.1], choosing estimators in this way gives an estimator that asymptotically has the same variance as the better of the ordinary and weighted least squares estimators.

6 Monte Carlo Simulations and Empirical Application

In this section, we present simulations studying the accuracy of the bootstrap approximations, as well as the efficiency of the convex-combination estimator in comparison with the ordinary and weighted least squares estimators. Simulations are given for univariate regression models in Section 6.1 and for multivariate models in Section 6.2. An empirical application is given in Section 6.3. We give the coverage and average length of bootstrap and asymptotic approximation confidence intervals. Because of the duality between intervals and testing, we omit simulations for tests. The tables presented compare the ordinary least squares estimator, the weighted least squares estimator, the estimator chosen between the ordinary and weighted estimators based on which has smaller sample variance, and the convex-combination estimator giving smallest sample variance (referred to as OLS, WLS, Min, and Optimal, respectively). Simulations are also given for the adaptive least squares (ALS) estimator. For this estimator, two methods of wild bootstrap-t intervals are given. The first recomputes the ALS estimator by performing a test for heteroskedasticity on each bootstrap sample (and is referred to as ALS1 in the tables) and the other reports the bootstrap-t interval of the estimator chosen by the test for heteroskedasticty (and is referred to as ALS2 in the tables).

Each of the covariance estimators given in Remark 3.1 can be used for computing standard errors. For covariance estimation, the HC2 and HC3 estimators outperform either the HC0 or HC1 estimators. The HC3 estimator may not always outperform the HC2 estimator, but is claimed in Flachaire (2005) to outperform the HC2 estimator in many situations. For this reason, in each of the simulations presented, the HC3 covariance estimator is used. Further simulations, which are omitted here, indicated that the performance of the bootstrap intervals are relatively insensitive to the choice of covariance estimator, but the HC3 estimator performed noticeably better for the asymptotic intervals than the other estimators. Intervals based on a t-approximation use 10,000 simulations, while bootstrap intervals use 10,000 simulations with 1,000 bootstrap samples. The bootstrap intervals presented are given by the wild bootstrap-t methods. Unless otherwise specified, the errors for the wild bootstrap distribution are generated using the F_2 (or Rademacher) distribution, which puts equal mass on ± 1 (as defined in Remark 3.2). In the bootstrap simulations, we scale the residuals (from the ordinary least squares estimator) by $1/\sqrt{1-h_i}$ when generating bootstrap samples, where the h_i are defined as in Remark 3.1. All confidence intervals are constructed with

a nominal coverage probability of 95%.

Throughout, the parametric family used to estimate the skedastic function is

$$v_\theta(x) := \exp\left(\theta_0 + \theta_1 \log|x_1| + \cdots + \theta_p \log|x_p|\right) \ ,$$

and $\hat{\theta}$ is found by the OLS solution to the regression problem

$$\log \max\left\{\hat{\epsilon}_i^2, \delta^2\right\} = \theta_0 + \theta_1 \log|x_1| + \cdots + \theta_p \log|x_p| + u_i$$

where u_i is the error term and $\delta := .1$. This method of estimating the skedastic function is also used in Romano and Wolf (2017). For the ALS estimator, the test for conditional heteroskedasiticity is the usual F-test of the hypothesis $H : \theta_1 = \cdots = \theta_p = 0$ at the 5% level.

6.1 Univariate models

Simulations are given using the model

$$y_i = \alpha + x_i\beta + \sqrt{v(x_i)}\varepsilon_i$$

where $x_i \sim U(1, 4)$ and ε_i are i.i.d. according to a distribution specified in several scenarios below. Several forms of the true skedastic function $v(\cdot)$ are used, and are specified in the tables. In each of the simulations, $(\alpha, \beta) = (0, 0)$ and a confidence interval is constructed for β.

Table 8.1 gives the empirical mean squared error when the errors, ε_i, are $N(0, 1)$. Table 8.2 gives the coverage of and average length of t intervals. To understand the effect of skewness of the error distributions, these simulations are repeated using exponential (with parameter one, centered to have mean zero) errors in Table 8.4 (with HC3 estimators).

Table 8.3 give the coverage and average length of wild bootstrap-t intervals when the errors are $N(0, 1)$. Simulations with exponential errors are given in Table 8.5. Table 8.6 repeats the simulations in Table 8.5, but instead uses the F_1 distribution (as defined in Remark 3.2) to generate the wild bootstrap error terms.

The empirical mean squared error of the weighted least squares estimator (Table 8.1) can be considerably smaller than that of the ordinary least squares estimator when the skedastic function is well modeled. When the family of skedastic functions is misspecified or there is conditional homoskedasticity, the weighted least squares may have worse mean squared error. While in several of the simulations, the empirical mean squared error of the weighted least squares estimator can be reduced by the ordinary least squares estimator, using the optimal combination, or the estimator with smallest estimated variance gives similar performance to the better of the ordinary and weighted least squares estimators. The adaptive least squares estimator has mean squared error that is close to the better of the ordinary and weighted least squares estimators, but can have somewhat larger mean squared error than the optimal combination estimator.

For normal errors, the asymptotic approximation intervals have coverage that is very close to the nominal level when using the ordinary least squares estimator. However, for each of the other estimators, the corresponding asymptotic intervals can have coverage that is noticeably under the nominal level (especially in small samples). Furthermore, coverage of the t intervals based on either the minimum variance or optimal convex-combination estimator is somewhat lower than the coverage of intervals based on either the ordinary or weighted least squares estimators. By comparison, the intervals using the wild bootstrap-t method have coverage that is closer to the nominal level than those based on an asymptotic approximation. For any estimator, the bootstrap intervals have comparable width to the corresponding t intervals.

In homoskedastic models, the size of the bootstrap-t intervals based on the convex-combination estimator are only very slightly wider than those given by the ordinary least squares estimator using the asymptotic approximation, and the intervals have comparable levels of coverage for each of the sample sizes studied. In the heteroskedastic models, the convex-combination estimator performs comparably to the weighted least squares estimator, even in small samples (e.g., n=20). By comparison, the adaptive least squares estimator gives intervals that tend to be somewhat wider than the weighted least squares estimator in small samples. In moderate and large samples, the adaptive least squares estimator performs comparably to the weighted least squares estimator. In each of the simulations, intervals based on the convex-combination estimator perform similarly to using the weighted least squares estimator in situations when this estimator is more efficient, but never perform noticeably worse than intervals based on the ordinary least squares estimator.

As with normal errors, when the errors follow an exponential distribution, the wild bootstrap-t intervals improve coverage over the asymptotic approximation intervals. However, even when using the bootstrap intervals, the coverage can be much below the nominal level for any of the estimators aside from the ordinary least squares estimator. In this setting, the performances of the optimal convex-combination estimator, and the adaptive least squares estimator are very similar.

Theoretical results, such as those given in Liu (1988), suggest that using the F_1 distribution may have better coverage than the F_2 distribution when the errors are skewed. The simulations indicate that even with skewed errors, the F_2 distribution has better small-sample performance. The findings here are in agreement with the simulation study provided in Davidson and Flachaire (2008). This paper asserts that "the F_2 distribution is never any worse behaved than the F_1 version, and is usually markedly better."

In the univariate setting with normally distributed errors, there is very little downside to using the optimal convex-combination estimator when compared with the ordinary least squares estimator, and this estimator often significantly improves efficiency. In small samples, the bootstrap intervals have coverage that is closer to the nominal level than the corresponding asymptotic approximation intervals. When the errors are very skewed, weighting can improve efficiency, and the bootstrap intervals again give better coverage, although the coverage can be much lower than the

nominal level. If the errors are severely skewed, it may not be worth weighting in very small sample sizes as the coverage for any of the estimators other than the ordinary least squares estimator can be severely below the nominal level.

6.2 Multivariate models

Simulations are given using the model

$$y_i = \alpha + x_{i,1}\beta_1 + x_{i,2}\beta_2 + x_{i,3}\beta_3 + \sqrt{v(x_i)}\varepsilon_i$$

where the $x_{i,j} \sim U(1,4)$ for $j = 1,2,3$ and $\varepsilon_i \sim N(0,1)$. Several forms of the true skedastic function $v(\cdot)$ are used, and are specified in the tables. Without loss of generality, the regression coefficients are all set to zero, and a confidence interval is constructed for β_1. In this section, simulations are given for homoskedastic models as well as heteroskedastic models using the following skedastic functions:

- $v_1(x) := \exp\left(2\log|x_1| + 2\log|x_2| + 2\log|x_3|\right)$
- $v_2(x) := \left(|x_1| + |x_2| + |x_3|\right)^2$
- $v_3(x) := \left(|x_1|^2 + |x_2|^2 + |x_3|^2\right)$
- $v_4(x) := \exp\left(\frac{2}{3}|x_1| + \frac{2}{3}|x_2| + \frac{2}{3}|x_3|\right)$

Table 8.8 gives the coverage and average length of t intervals and Table 8.7 gives the coverage and average length of wild bootstrap-t intervals.

These simulations demonstrate that intervals based on the weighted least squares estimator, or the optimal convex-combination estimator found using an asymptotic approximation have coverage that is below the nominal level. In small samples ($n = 20$), the coverage of the intervals based on the bootstrap is closer to the nominal level than the asymptotic approximation intervals, although the coverage can be somewhat below the nominal level. In moderate sample sizes ($n = 50$), the coverage of the bootstrap intervals is almost exactly at the nominal level in each of the examples, whereas the asymptotic intervals can still have coverage that is noticeably below the nominal level. In each of the examples, the optimal convex-combination estimator performs comparably to the better of the weighted and ordinary least squares estimators.

In small samples, there appears to be little improvement in efficiency from weighting, but the coverage for each of the weighted estimators tends to be somewhat lower than the coverage for the intervals based on the ordinary least squares estimator. Therefore, in small sample sizes ($n = 20$), it may be better to use the ordinary least squares estimator. In more moderate samples (n=50), there can be substantial improvements in efficiency from weighting. The optimal convex-combination estimator performs comparably to the better of the ordinary and weighted least squares estimators.

In comparison, the adaptive least squares can be more efficient than the ordinary least squares estimator, but is often less efficient than either the convex-combination estimator, or the weighted least squares estimator. Therefore, if the sample size is relatively small, it may be best to report the asymptotic intervals from the ordinary least squares estimator. In moderate and large sample size, the optimal convex-combination estimator gives nearly best performance in each of the simulations. Especially in moderate sample sizes (n=50), the coverage of the intervals based on this interval is improved by using the bootstrap.

6.3 Empirical Example

The dataset under consideration is a sample of 506 observations taken from the Boston area in 1970. Five of the included variables are:

log (price):	log of median house price in US dollars
log(nox)	log of nitrogen oxide in the air in ppm
log(dist)	log of weighted distance from employment centers in miles
rooms	average number of rooms per house
stratio	average student-teacher ratio

The response variable is log(price), and the four remaining variables are the explanatory variables. The family of skedastic functions used to estimate the true skedastic function, as well as the method of estimating the parameter, is that used in Section 6.2 but extended to have two additional predictors. Table 8.9 gives the estimates of the coefficients for each of the predictors. Table 8.10 gives the corresponding confidence intervals. Table 8.11 gives the lengths of the intervals in Table 8.10.

The estimated coefficient of *stratio* from the optimal convex-combination estimator is between the ordinary and weighted least squares estimator. For this coefficient, the interval is narrower for the convex-combination estimator than either the ordinary or weighted least squares estimator (and also the adaptive least squares estimator which agrees with the weighted least squares estimator). For the remaining variables, the estimated coefficients using the optimal convex-combination estimator are identical to those using the weighted least squares estimator which produces narrower intervals than the ordinary least squares estimator. For these coefficients, the intervals from the convex-combination estimator are nearly identical to those from the weighted least squares estimator. This example confirms that for large sample sizes, the optimal convex-combination estimator produces intervals that are nearly identical to the narrower of the intervals given by the weighted and ordinary least squares estimators, if not even narrower.

7 Conclusion

Making some attempt to model the skedastic function and using a weighted estimator can result in large gains in efficiency when compared with inference based on ordinary least squares estimators. Still, there are some shortcomings to basing inference on a weighted least squares estimator, with heteroskedasticity-consistent standard errors (which are valid when the skedastic function is not consistently estimated), and using an asymptotic approximation to the sampling distribution. Simulations demonstrate that asymptotic approximations can give poor small sample performance, yielding confidence intervals with coverage below the nominal level, or tests with type I error rates that can be larger than the nominal level. Furthermore, a badly estimated skedastic function can result in an estimator that is less efficient than simply using the ordinary least squares estimator irrespective of the sample size.

In this paper, we propose an estimator that is a convex-combination between the ordinary and weighted least squares estimators. The convex-combination estimator takes advantage of weighting when weighting provides improvement in efficiency, and performs comparably to the OLS otherwise. There is little downside, even in homoscedastic models, to using the convex-combination estimator rather than the OLS estimator. But in circumstances when the WLS estimator is advantageous, the convex-combination estimator has comparable performance to the WLS estimator. Simulations confirm that the convex-combination estimator performs similarly to the better of the WLS and OLS estimators. In contrast, the adaptive least squares estimator may not realize all of the efficiency gains to be had by weighting, especially in small and moderate sample sizes.

For either the weighted least squares estimator or the convex-combination estimator, inference based on asymptotic approximations to the sampling distributions can have poor performance in small or even moderate sample sizes. This paper established consistency of the pairs and wild bootstrap for both of these estimators. Simulations demonstrated that in small or moderate samples, using the bootstrap approximations has improved coverage for confidence intervals. Of course, the bootstrap often has higher-order accuracy when compared with asymptotic approximations as discussed in Hall (1992). Proving improvements in accuracy from the bootstrap in our application is an open question, but would require accounting for the data-driven choice of weights, and is beyond the scope of the paper. Inference using the convex-combination estimator bridges the gap between the ordinary and weighted least squares estimator. Unless the sample size is very small relative to the number of coefficients under consideration, in which case weighting may only provide relatively modest benefits, the convex-combination estimator is never noticeably worse than the ordinary least squares estimator, and potentially much better. In small and moderate samples, using a bootstrap approximation to the sampling distribution leads to more reliable inference.

8 Tables

	OLS	WLS	Min	Optimal	ALS
n = 20, $v(x) = 1$	0.0754	0.0838	0.0795	0.0794	0.0764
n = 50, $v(x) = 1$	0.0284	0.0297	0.0294	0.0292	0.0282
n = 100, $v(x) = 1$	0.0136	0.0140	0.0140	0.0138	0.0137
n = 20, $v(x) = x^2$	0.5611	0.4550	0.4824	0.4775	0.5291
n = 50, $v(x) = x^2$	0.2107	0.1555	0.1637	0.1627	0.1787
n = 100, $v(x) = x^2$	0.0511	0.0352	0.0363	0.0360	0.0745
n = 20, $v(x) = \log(x)^2$	0.0654	0.0457	0.0483	0.0487	0.0582
n = 50, $v(x) = \log(x)^2$	0.0249	0.0137	0.0138	0.0146	0.0152
n = 100, $v(x) = \log(x)^2$	0.0123	0.0063	0.0062	0.0065	0.0063
n = 20, $v(x) = 4\exp(.02x + .02x^2)$	0.3613	0.4088	0.3943	0.3816	0.3651
n = 50, $v(x) = 4\exp(.02x + .02x^2)$	0.1368	0.1450	0.1390	0.1405	0.1392
n = 100, $v(x) = 4\exp(.02x + .02x^2)$	0.0667	0.0686	0.0682	0.0677	0.0678

Table 8.1: Empirical mean squared error of estimators of β.

		OLS	WLS	Min	Optimal	ALS
n = 20, $v(x) = 1$	Coverage	0.9507	0.9353	0.9340	0.9338	0.9477
	Length	1.1950	1.1608	1.1341	1.1301	1.1866
n = 50, $v(x) = 1$	Coverage	0.9491	0.9423	0.9412	0.9411	0.9483
	Length	0.6805	0.6755	0.6669	0.6659	0.6789
n = 100, $v(x) = 1$	Coverage	0.9500	0.9449	0.9457	0.9463	0.9479
	Length	0.4661	0.4646	0.4616	0.4612	0.4656
n = 20, $v(x) = x^2$	Coverage	0.9476	0.9425	0.9355	0.9349	0.9401
	Length	3.2361	2.8017	2.7418	2.7117	3.0106
n = 50, $v(x) = x^2$	Coverage	0.9438	0.9433	0.9380	0.9359	0.9380
	Length	1.8600	1.5711	1.5637	1.5500	1.6275
n = 100, $v(x) = x^2$	Coverage	0.9465	0.9482	0.9469	0.9458	0.9475
	Length	1.2761	1.0641	1.0634	1.0574	1.0817
n = 20, $v(x) = \log(x)^2$	Coverage	0.9463	0.9495	0.9406	0.9388	0.9421
	Length	1.1017	0.8774	0.8687	0.8595	0.9496
n = 50, $v(x) = \log(x)^2$	Coverage	0.9461	0.9516	0.9498	0.9466	0.9443
	Length	0.6375	0.4706	0.4704	0.4675	0.4746
n = 100, $v(x) = \log(x)^2$	Coverage	0.9465	0.9498	0.9496	0.9477	0.9516
	Length	0.4379	0.3134	0.3134	0.3125	0.3130
n = 20, $v(x) = 4\exp(.02x + .02x^2)$	Coverage	0.9548	0.9388	0.9358	0.9368	0.9470
	Length	2.6677	2.6016	2.5252	2.5134	2.6386
n = 50, $v(x) = 4\exp(.02x + .02x^2)$	Coverage	0.9512	0.9431	0.9435	0.9437	0.9516
	Length	1.5151	1.5042	1.4807	1.4778	1.5099
n = 100, $v(x) = 4\exp(.02x + .02x^2)$	Coverage	0.9516	0.9497	0.9484	0.9492	0.9529
	Length	1.0375	1.0338	1.0245	1.0234	1.0351

Table 8.2: Coverage and average length of confidence intervals for β based on an asymptotic approximation using HC3 standard errors.

		OLS	WLS	Min	Optimal	ALS1	ALS2
	Coverage	0.9463	0.9447	0.9439	0.9438	0.9448	0.9435
n = 20, $v(x) = 1$	Length	1.1935	1.2535	1.2298	1.2262	1.2157	1.1952
	Coverage	0.9503	0.9484	0.9514	0.9506	0.9486	0.9471
n = 50, $v(x) = 1$	Length	0.6775	0.6967	0.6889	0.6969	0.6804	0.6748
	Coverage	0.9476	0.9479	0.9481	0.9477	0.9485	0.9482
n = 100, $v(x) = 1$	Length	0.4640	0.4706	0.4677	0.4671	0.0.4699	0.4697
	Coverage	0.9432	0.9470	0.9447	0.9449	0.9425	0.9403
n = 20, $v(x) = x^2$	Length	3.3621	3.0161	3.0451	3.0251	3.32317	3.1048
	Coverage	0.9483	0.9478	0.9459	0.9465	0.9471	0.9414
n = 50, $v(x) = x^2$	Length	1.8844	1.5971	1.6253	1.6144	1.6889	1.6472
	Coverage	0.9475	0.9515	0.9512	0.9527	0.9504	0.0.9511
n = 100, $v(x) = x^2$	Length	1.2874	1.0733	1.0791	1.0782	1.0698	1.0832
	Coverage	0.9417	0.9510	0.9487	0.9494	0.9418	0.9353
n = 20, $v(x) = \log(x)^2$	Length	1.1823	0.9407	0.9718	0.9648	1.0645	0.9703
	Coverage	0.9487	0.9516	0.9521	0.9508	0.9484	0.9436
n = 50, $v(x) = \log(x)^2$	Length	0.6505	0.4704	0.4774	0.4793	0.4828	0.4739
	Coverage	0.9490	0.9485	0.9497	0.9486	0.9488	0.9492
n =100, $v(x) = \log(x)^2$	Length	0.4424	0.3116	0.3116	0.3140	0.3126	0.3126
	Coverage	0.9445	0.9420	0.9431	0.9428	0.9456	0.9439
n = 20, $v(x) = 4\exp(.02x + .02x^2)$	Length	2.6782	2.8347	2.7696	2.7579	2.7275	2.6663
	Coverage	0.9474	0.9484	0.9461	0.9485	0.9450	0.9440
n = 50, $v(x) = 4\exp(.02x + .02x^2)$	Length	1.5091	1.5522	1.5309	1.5256	1.5183	1.5050
	Coverage	0.9526	0.9492	0.9507	0.9513	0.9511	0.9504
n = 100, $v(x) = 4\exp(.02x + .02x^2)$	Length	1.0336	1.0459	1.0384	1.0369	1.0372	1.0364

Table 8.3: Coverage and average length of confidence intervals for β based on the wild bootstrap-t method with HC3 covariance estimates.

		OLS	WLS	Min	Optimal	ALS
n = 20, $v(x) = 1$	Coverage	0.9636	0.9274	0.9280	0.9266	0.9422
	Length	1.1500	1.0756	1.0464	1.0410	1.1127
n = 20, $v(x) = x^2$	Coverage	0.9185	0.9101	0.9035	0.9013	0.9121
	Length	3.0413	2.5939	2.5196	2.4868	2.7363
n = 20, $v(x) = \log(x)^2$	Coverage	0.9099	0.9058	0.8992	0.8981	0.9046
	Length	1.0341	0.8317	0.8161	0.8058	0.8750
n = 20, $v(x) = 4\exp(.02x + .02x^2)$	Coverage	0.9605	0.9266	0.9247	0.9240	0.9426
	Length	2.5280	2.3657	2.2899	2.2742	2.4477

Table 8.4: Coverage and average length of confidence intervals for β based on the asymptotic approximation using the HC3 covariance estimator with exponential errors.

		OLS	WLS	Min	Optimal	ALS1	ALS2
n = 20, $v(x) = 1$	Coverage	0.9557	0.9292	0.9322	0.9348	0.9388	0.9355
	Length	1.1364	1.1334	1.1167	1.1112	1.1181	1.0967
n = 20, $v(x) = x^2$	Coverage	0.9316	0.9355	0.9221	0.9201	0.9210	0.9201
	Length	3.0863	2.7799	2.7749	2.7494	2.8986	2.7847
n = 20, $v(x) = \log(x)^2$	Coverage	0.9171	0.9238	0.9040	0.9070	0.9045	0.8998
	Length	1.1605	0.9568	0.9093	0.9009	0.9294	0.8847
n = 20, $v(x) = 4\exp(.02x + .02x^2)$	Coverage	0.9680	0.9429	0.9357	0.9363	0.9392	0.9351
	Length	2.7516	2.7648	2.5510	2.5103	2.5088	2.4565

Table 8.5: Coverage and average length of wild bootstrap-t confidence intervals for β using the HC3 covariance estimator with exponential errors.

		OLS	WLS	Min	Optimal	ALS1	ALS2
n = 20, $v(x) = 1$	Coverage	0.9193	0.8849	0.8885	0.8890	0.9008	0.8948
	Length	1.0042	1.0078	0.9890	0.9838	0.9976	0.9711
n = 20, $v(x) = x^2$	Coverage	0.8851	0.8966	0.8929	0.8943	0.8882	0.8747
	Length	2.6845	2.4739	2.4387	2.4162	2.5394	2.4354
n = 20, $v(x) = \log(x)^2$	Coverage	0.8691	0.8939	0.8864	0.8860	0.8828	0.8666
	Length	0.9151	0.7735	0.7753	0.7676	0.8106	0.7730
n = 20, $v(x) = 4\exp(.02x + .02x^2)$	Coverage	0.9166	0.8857	0.8878	0.8873	0.8963	0.8922
	Length	2.2245	2.2604	2.2023	2.1942	2.2201	2.1609

Table 8.6: Coverage and average length of wild bootstrap-t (generated using Mammen's error distribution) confidence intervals for β using the HC3 covariance estimator with exponential errors.

		OLS	WLS	Min	Optimal	ALS1	ALS2
	Coverage	0.9420	0.9376	0.9381	0.9360	0.9418	0.9383
n = 20, $v(x) = 1$	Length	1.3119	1.4537	1.4140	1.4017	1.3421	1.3106
	Coverage	0.9517	0.9473	0.9479	0.9486	0.9473	0.9448
n = 50, $v(x) = 1$	Length	0.6960	0.7470	0.7280	0.7222	0.6983	0.6904
	Coverage	0.9496	0.9510	0.9499	0.9471	0.9492	0.9476
n = 100, $v(x) = 1$	Length	0.4696	0.4889	0.4813	0.4773	0.4726	0.4698
	Coverage	0.9467	0.9496	0.9494	0.9490	0.9434	0.9350
n = 20, $v(x) = v_1(x)$	Length	24.4902	22.7113	22.4493	22.1392	23.795	22.6830
	Coverage	0.9492	0.9533	0.9547	0.9545	0.9517	0.9349
n = 50, $v(x) = v_1(x)$	Length	13.1570	9.3530	9.5490	9.3708	9.4461	10.4717
	Coverage	0.9514	0.9582	0.9573	0.9568	0.9569	0.9553
n = 100, $v(x) = v_1(x)$	Length	8.9963	5.4756	5.5501	5.4863	5.5503	5.4984
	Coverage	0.9424	0.9404	0.9391	0.9398	0.9377	0.9337
n = 20, $v(x) = v_2(x)$	Length	10.0941	11.2374	10.7655	10.6653	10.2414	9.9523
	Coverage	0.9517	0.9528	0.9510	0.9506	0.9480	0.9458
n = 50, $v(x) = v_2(x)$	Length	5.3449	5.5378	5.4130	5.3408	5.3900	5.2702
	Coverage	0.9512	0.9494	0.9504	0.9485	0.9481	0.9430
n =100, $v(x) = v_2(x)$	Length	3.6078	3.5658	3.5518	3.4946	3.6213	3.5610
	Coverage	0.9373	0.9349	0.9370	0.9369	0.9382	0.9377
n = 20, $v(x) = v_3(x)$	Length	6.0691	6.7292	6.4700	6.4147	6.0984	6.0871
	Coverage	0.9487	0.9454	0.9465	0.9488	0.9482	0.9432
n = 50, $v(x) = v_3(x)$	Length	3.2075	3.3498	3.2673	3.2264	3.2659	3.1943
	Coverage	0.9492	0.9501	0.9493	0.9484	0.9470	0.9450
n = 100, $v(x) = v_3(x)$	Length	2.1843	2.1817	2.1620	2.1316	2.1957	2.1632
	Coverage	0.9471	0.9434	0.9461	0.9450	0.9440	0.9376
n = 20, $v(x) = v_4(x)$	Length	20.8139	20.5321	20.0122	19.7953	20.8253	19.9887
	Coverage	0.9504	0.9489	0.9501	0.9490	0.9447	0.9338
n = 50, $v(x) = v_4(x)$	Length	11.1697	9.1714	9.2610	9.1012	10.1090	9.6712
	Coverage	0.9516	0.9496	0.9501	0.9492	0.9552	0.9505
n = 100, $v(x) = v_4(x)$	Length	7.6657	5.6528	5.7152	5.6237	5.8274	5.7378

Table 8.7: Coverage and average length of confidence intervals for β_1 based on the wild bootstrap-t method.

		OLS	WLS	Min	Optimal	ALS
	Coverage	0.9665	0.9264	0.9284	0.9280	0.9586
n = 20, $v(x) = 1$	Length	1.3649	1.2433	1.2063	1.1962	1.3529
	Coverage	0.9547	0.9281	0.9303	0.9305	0.9525
n = 50, $v(x) = 1$	Length	0.7127	0.6927	0.6779	0.6747	0.7072
	Coverage	0.9501	0.9401	0.9384	0.9404	0.9510
n = 100, $v(x) = 1$	Length	0.4767	0.4715	0.4654	0.4641	0.4759
	Coverage	0.9675	0.9482	0.9398	0.9387	0.9533
n = 20, $v(x) = v_1(x)$	Length	24.8500	19.1977	18.5013	18.0865	22.7933
	Coverage	0.9572	0.9551	0.9473	0.9460	0.9440
n = 50, $v(x) = v_1(x)$	Length	13.3704	8.8556	8.7810	8.5842	9.6507
	Coverage	0.9535	0.9604	0.9577	0.9544	0.9588
n = 100, $v(x) = v_1(x)$	Length	9.0588	5.4373	5.4267	5.3342	5.4555
	Coverage	0.9607	0.9234	0.9215	0.9195	0.9562
n = 20, $v(x) = v_2(x)$	Length	10.4320	9.5058	9.0461	8.9017	10.2254
	Coverage	0.9541	0.9362	0.9346	0.9334	0.9469
n = 50, $v(x) = v_2(x)$	Length	5.4633	5.1493	4.9956	4.9407	5.3603
	Coverage	0.9532	0.9358	0.9364	0.9375	0.9431
n =100, $v(x) = v_2(x)$	Length	3.6640	3.4356	3.3935	3.3643	3.5411
	Coverage	0.9621	0.9238	0.9209	0.9208	0.9566
n = 20, $v(x) = v_3(x)$	Length	6.2806	5.6983	5.4457	5.3680	6.1906
	Coverage	0.9562	0.9321	0.9303	0.9318	0.9473
n = 50, $v(x) = v_3(x)$	Length	3.2997	3.1127	3.0222	2.9923	3.2495
	Coverage	0.9551	0.9411	0.9407	0.9412	0.9475
n = 100, $v(x) = v_3(x)$	Length	2.2141	2.0956	2.0678	2.0522	2.1572
	Coverage	0.9645	0.9324	0.9248	0.9245	0.9550
n = 20, $v(x) = v_4(x)$	Length	21.2127	17.1633	16.5004	16.1612	20.2824
	Coverage	0.9537	0.9454	0.9387	0.9379	0.9386
n = 50, $v(x) = v_4(x)$	Length	11.4171	8.5834	8.4802	8.3146	9.4878
	Coverage	0.9503	0.9485	0.9452	0.9427	0.9497
n = 100, $v(x) = v_4(x)$	Length	7.7106	5.5034	5.4869	5.4028	5.6219

Table 8.8: Coverage and average length of confidence intervals for β_1 based on an asymptotic approximation.

Coefficient	OLS	WLS	Min	Optimal
Constant	11.0838	10.1952	10.1952	10.1952
log(nox)	−0.9535	−0.7934	−0.7934	−0.7934
log(dist)	−0.1343	−0.1265	−0.1265	−0.1265
rooms	0.2545	0.3065	0.3065	0.3065
stratio	−0.0525	−0.0367	−0.0525	−0.0451

Table 8.9: Estimated coefficients for each predictor.

	Constant	log(nox)	log(dist)	rooms	stratio
OLS	(10.3236 , 11.8411)	(−1.2068 , −0.7010)	(−0.2406, −0.0260)	(0.2047, 0.3046)	(−0.0614, −0.0433)
WLS	(9.6224, 10.7555)	(−0.9976, −0.5859)	(−0.2007 , −0.0526)	(0.2741, 0.3396)	(−0.0460 , −0.0274)
Min	(9.6079, 10.7598)	(−0.9960 , −.5872)	(−0.1998 , −0.0527)	(0.2734, 0.3396)	(−0.0621, −0.0430)
Opt	(9.6336, 10.7702)	(−0.9970, −0.5924)	(−0.2001, −0.0537)	(0.2732 , 0.3399)	(−0.0541 , −0.0361)
ALS1	(9.6096, 10.7613)	(−0.9969, −0.5810)	(−0.1996 , −0.0527)	(0.2732 , 0.3400)	(−0.0459 , −0.0272)
ALS2	(9.6224, 10.7555)	(−0.9976, −0.5859)	(−0.2007 , −0.0526)	(0.2741, 0.3396)	(−0.0460 , −0.0274)

Table 8.10: Confidence intervals for each predictor.

	Constant	log(nox)	log(dist)	rooms	stratio
OLS	1.5175	0.5058	0.2146	0.0999	0.0181
WLS	1.1331 (0.7467)	0.4117 (0.8140)	0.1481 (0.6901)	0.0655 (0.6557)	0.0186 (1.0276)
Min	1.1519 (0.7591)	0.4088 (0.8082)	0.1471 (0.6855)	0.0662 (0.6627)	0.0191 (1.0552)
Opt	1.1366 (0.7490)	0.4046 (0.7999)	0.1464 (0.6822)	0.0667 (0.6677)	0.0180 (0.9945)
ALS1	1.1517 (0.7589)	0.4159 (0.8223)	0.1469 (0.6845)	0.0668 (0.6687)	0.0187 (1.0331)
ALS2	1.1331 (0.7467)	0.4117 (0.8140)	0.1481 (0.6901)	0.0655 (0.6557)	0.0186 (1.0276)

Table 8.11: Length of intervals for each predictor and the length expressed as a ratio of the length of the OLS intervals in parenthesis.

9 Appendix

Proof of Theorem 3.1. For a fixed function $w(\cdot)$, define $W \coloneqq \operatorname{diag}\{w(x_1), ..., w(x_n)\}$ and

$$\hat{\beta}_W \coloneqq (X^\top W^{-1} X)^{-1} X^\top W^{-1} Y \ .$$

If the skedastic function is estimated from a family $\{v_\theta\}$ by $v_{\hat{\theta}}$, the weighted least squares estimator is given by by

$$\hat{\beta}_{\text{WLS}} \coloneqq (X^\top V_{\hat{\theta}}^{-1} X)^{-1} X^\top V_{\hat{\theta}}^{-1} Y$$

where $V_\theta \coloneqq \operatorname{diag}\{v_\theta(x_1), ..., v_\theta(x_n)\}$. We would like to show that the bootstrap distribution $\sqrt{n}\big(\hat{\beta}^*_{WLS} - \hat{\beta}_{\text{WLS}}\big)$ (conditional on the data) consistently approximates the sampling distribution of $\sqrt{n}\big(\hat{\beta}_{\text{WLS}} - \beta\big)$. To do this, we will first show that the distribution of $\sqrt{n}\big(\hat{\beta}^*_W - \hat{\beta}_W\big)$ consistently approximates the distribution of $\sqrt{n}\big(\hat{\beta}_W - \beta\big)$ for a fixed W (satisfying some regularity conditions). We will then show that $\sqrt{n}\big(\hat{\beta}^*_{WLS} - \hat{\beta}_{\text{WLS}}\big) - \sqrt{n}\big(\hat{\beta}^*_W - \hat{\beta}_W\big)$ converges in conditional probability to zero for $W = V_{\theta_0}$, assuming that the estimate $\hat{\theta}^*$ of the variance parameter is conditionally consistent for some fixed θ_0. That is, the proof of Theorem 3.1 will rely on Lemmas 9.1 and 9.2 which are stated below. ■

Lemma 9.1. *Suppose that $(x_1, y_1), ..., (x_n, y_n)$ are i.i.d. satisfying assumptions (A1)–(A6). Suppose that $w : \mathbb{R}^d \to \mathbb{R}^+$ is a fixed and known function (although not necessarily the true skedastic function) and satisfies*

$$\mathbb{E}\left(\frac{\left(y_i^2 + \sum_{j=1}^p x_{i,j}^2\right)^2}{w^2(x_i)}\right) < \infty$$

Define $W \coloneqq diag(w(x_1), ..., w(x_n))$, and let $\hat{\beta}_W \coloneqq (X^\top W^{-1} X)^{-1} X^\top W^{-1} Y$. Then, for almost all sample sequences, the conditional law of $\sqrt{n}\big(\hat{\beta}^*_W - \hat{\beta}_W\big)$ converges weakly to the normal distribution with mean 0 and variance $\Omega_{1/w}^{-1} \Omega_{v/w^2} \Omega_{1/w}^{-1}$.

Proof of Lemma 9.1 using the pairs bootstrap. Let C_P be the set of sequences $\{P_n\}$ such that

(B1) P_n converges weakly to P (the distribution of (x_i, y_i)).

(B2) $\beta_W(P_n) \coloneqq \left(\int \frac{1}{w(x)} x x^\top dP_n\right)^{-1} \cdot \int \frac{1}{w(x)} x y dP_n \to \beta$.

(B3) $\int \frac{1}{w(x)} x x^\top dP_n \to \Omega_{1/w}$.

(B4) $\int \big(1/w(x) x^\top (y - x\beta_W(P_n))\big)^\top \big(1/w(x) x^\top (y - x\beta_W(P_n))\big) dP_n \to \Omega_{v/w^2}$.

To prove the lemma, we will first show that the distribution of $\sqrt{n}\big(\hat{\beta}_W - \beta_W(P_n)\big)$ under P_n converges weakly to the normal distribution with mean 0 and variance $\Omega_{1/w}^{-1} \Omega_{v/w^2} \Omega_{1/w}^{-1}$ whenever $\{P_n\} \in C_p$, and then show that the empirical distribution is in C_p almost surely.

Let $(x_{n,i}, y_{n,i})$, $i = 1, ..., n$ be independent and identically distributed according to P_n such that $\{P_n\} \in C_P$.

Define residuals $\varepsilon_{n,i} \coloneqq Y_{n,i} - X_{n,i}\beta_W(P_n)$ so that

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}_W - \beta_W(P_n)\right) &= \sqrt{n}\left(X_n^\top W^{-1} X_n\right)^{-1} X_n^\top W^{-1}\left(\varepsilon_n + X_n\beta_W(P_n)\right) - \beta_W(P_n) \\ &= \left(\frac{1}{n} X_n^\top W^{-1} X_n\right)^{-1} \sqrt{n} X_n^\top W^{-1}\varepsilon_n \ .\end{aligned}$$

It follows immediately from the assumptions that

$$\left(\frac{1}{n} X_n^\top W^{-1} X_n\right)^{-1} \xrightarrow{P} \Omega_{1/w}^{-1} \ ,$$

and we have the desired asymptotic normal distribution if we can show

$$\sqrt{n} X_n^\top W^{-1}\varepsilon_n \xrightarrow{d} N(0, \Omega_{v/w^2}) \ .$$

We will first consider the case of $x_i \in \mathbb{R}$. Because

$$\int x_{n,i}^\top \frac{1}{w(x_{n,i})}(y_{n,i} - x_{n,i}\beta_W(P_n)) dP_n = 0 \ ,$$

and

$$\int x_{n,i}^\top x_{n,i} \frac{1}{w^2(x_{n,i})}\varepsilon_{n,i}^2 dP_n \to \Omega_{v/w^2} \ ,$$

the asymptotic normality follows from the Lindeberg-Feller Central Limit Theorem if we can verify that

$$\mathbb{E}\left(x_{n,1}^2 \frac{1}{w^2(x_{n,1})}\varepsilon_{n,1}^2 \quad \left\{x_{n,1}^2 \frac{1}{w^2(x_{n,1})}\varepsilon_{n,1}^2 > n\delta\right\}\right) \to 0$$

for all $\delta > 0$, where $\{\cdot\}$ denotes the indicator function of a set. Since $\beta_W(P_n) \to \beta$ and $(x_{n,i}, y_{n,i}) \xrightarrow{d} (X, Y) \sim P$,

$$x_{n,1} \frac{1}{w(x_{n,1})}\varepsilon_{n,1} \xrightarrow{d} \frac{X}{w(X)}(Y - X\beta) = \frac{X}{w(X)}\varepsilon \ .$$

By assumption (B4), we also have that the second moments converge in addition to the convergence in probability. Therefore, for any fixed γ that is a continuity point of the distribution of $X\varepsilon/w(X)$ and $n > \gamma/\delta$, we have that

$$\begin{aligned}\mathbb{E}\left(x_{n,1}^2 \frac{1}{w^2(x_{n,1})}\varepsilon_{n,1}^2 \quad \left\{x_{n,1}^2 \frac{1}{w^2(x_{n,1})}\varepsilon_{n,1}^2 > n\delta\right\}\right) &\leq E\left(x_{n,1}^2 \frac{1}{w^2(x_{n,1})}\varepsilon_{n,1}^2 \quad \left\{x_{n,1}^2 \frac{1}{w^2(x_{n,1})}\varepsilon_{n,1}^2 > \gamma\right\}\right) \\ &\to E\left(X^2 \frac{1}{w^2(X)}\varepsilon^2 \quad \left\{X^2 \frac{1}{w^2(X)}\varepsilon^2 > \gamma\right\}\right) \ .\end{aligned}$$

The Lindeberg-Feller condition is satisfied, since the right-hand side of this equation can be made arbitrarily small by choosing γ sufficiently large. The multivariate case follows analogously using the Cramér-Wold device. For any vector of constants, $C \in \mathbb{R}^p$, we must show

$$\sum_{i=1}^n \frac{\varepsilon_{n,i}}{w(x_{n,i})} x_{n,i} C \xrightarrow{d} N(0, C^\top \Omega_{v/w^2} C) \ .$$

This convergence follows from the Lindeberg-Feller CLT if

$$\mathbb{E}\left(\left(\frac{\varepsilon_{n,i}}{w(x_{n,i})}x_{n,i}C\right)^2 \quad \left\{\left(\frac{\varepsilon_{n,i}}{w(x_{n,i})}x_{n,i}C\right)^2 > n\delta\right\}\right) \to 0$$

for all $\delta > 0$. This convergence holds by the same argument as in the one-dimensional case given above. It is easily seen that the empirical distribution functions $\hat{P}_n$ are almost surely in C_P, and the result of the theorem follows. ■

Proof of Lemma 9.1 using the wild bootstrap. Let S be the set of sequences $\{x_i, y_i\}$ satisfying the following conditions:

(S1) $\hat{\beta}_W \to \beta$,

(S2) $\hat{\Omega}_{1/w} \to \Omega_{1/w}$,

(S3) $\hat{\Omega}_{v/w^2} \to \Omega_{v/w^2}$, and

(S4) $\sqrt{n}\left(\hat{\beta}_{\text{WLS}} - \hat{\beta}_W\right) \to 0$.

Write

$$\sqrt{n}\left(\hat{\beta}^*_W - \hat{\beta}_W\right) = \sqrt{n}\left(X_n^\top W^{-1} X_n\right)^{-1} X_n^\top W^{-1}\hat{\varepsilon}^* + \sqrt{n}\left(\hat{\beta}_{\text{WLS}} - \hat{\beta}_W\right) \ .$$

On S, $\left(\frac{1}{n}X_n^\top W^{-1}X_n\right)^{-1} \to \Omega_{1/w}$, and $\sqrt{n}\left(\hat{\beta}_{\text{WLS}} - \hat{\beta}_W\right) \to 0$. Thus, to show the desired asymptotic normality, it suffices to show that, on S, $W^{-1}\hat{\varepsilon}^* \xrightarrow{d} N(0, \Omega_{v/w^2})$ conditionally on the $x's$ and $y's$. This convergence holds using the Cramér-Wold device, since for each vector $c \in \mathbb{R}^p$,

$$c^\top X_n^\top W^{-1}\hat{\varepsilon}^* = \sum x_i c \frac{1}{w(x_i)}\hat{\varepsilon}^*$$

which is asymptotically normal with mean zero and variance $c^\top \Omega_{v/w^2} c$ by the Lindeberg-Feller Central Limit Theorem which is applicable because condition $(S3)$ holds.
The conditions specified by the set S do not hold almost surely, but they do hold in probability. By the Almost Sure Representation Theorem, there exist versions of the X's and Y's such that S holds almost surely. It follows that the asymptotic normality of the wild bootstrap distribution holds in probability. ■

Lemma 9.2. *Suppose that $\hat{\theta}^*$ is consistent for θ_0, in the sense that $n^{1/4}(\hat{\theta}^* - \theta_0)$ converges in conditional probability to zero. Suppose that $\hat{\beta}_{WLS} := (X^\top V_{\hat{\theta}}^{-1} X)^{-1} X^\top V_{\hat{\theta}}^{-1} Y$ and $v_{\theta_0} =: w$ so that $W := diag(v_{\theta_0}(X_1), ..., v_{\theta_0}(X_n))$. Under the assumptions of Theorem 3.1,*

$$\sqrt{n}(\hat{\beta}^*_{WLS} - \hat{\beta}_{WLS}) - \sqrt{n}(\hat{\beta}^*_W - \hat{\beta}_W) \xrightarrow{P} 0$$

in probability.

Proof of Lemma 9.2 using the pairs bootstrap. Let C_P be the set of sequences $\{P_n\}$ that satisfy the following conditions:

(C1) P_n converges weakly to P

(C2) $\int \frac{1}{w(x)} xx^\top dP_n \to \Omega_{1/w}$

(C3) $\int \left(1/w(x)x^\top(y - x\beta_W(P_n)\right)^\top \left(1/w(x)x^\top(y - x\beta_W(P_n)\right) dP_n \to \Omega_{v/w^2}$

(C4) $n^{1/4}\left(\beta_W(P_n) - \beta(P_n)\right) \to 0$

(C5) $n^{1/4}\mathbb{E}_{P_n}\left(x_i(y - x\beta(P_n))r_{\theta_{0,l}}(x)\right) \to 0$ for each $i = 1, ..., p$, $l = 1, ..., d$

(C6) $\mathbb{E}_{P_n}\left|x_i\varepsilon r_{\theta_{0,l}}(x)\right|^2 \to \mathbb{E}_P(|x_i\varepsilon r_{\theta_0,l}(x)|^2)$ for each $i = 1, ..., p$, $l = 1, ..., d$

(C7) $\mathbb{E}_{P_n}|x_i\varepsilon s_{\theta_0}(x)|^2 \to \mathbb{E}_P(|x_i\varepsilon s_{\theta_0}(x)|^2)$ for each $i = 1, ..., p$, $l = 1, ..., d$

(C8) $n^{1/4}\big(\hat{\theta} - \theta_0\big)$ converges in P_n-probability to zero

Suppose that $(x_{n,i}, y_{n,i})$, $i = 1, ..., n$ are i.i.d. according to P_n where $\{P_n\}$ is any sequence in C_P.

Define the residuals

$$\varepsilon_{\hat{W},n,i} := y_{n,i} - x_{n,i}\beta_{\hat{W}}(P_n) \ ,$$

$$\varepsilon_{n,i} := y_{n,i} - x_{n,i}\beta(P_n) \ ,$$

and

$$\varepsilon_{W,n,i} := y_{n,i} - x_{n,i}\beta_W(P_n)$$

where

$$\beta_{\hat{W}}(P_n) := \left(\int \frac{1}{v_{\hat{\theta}}(x)} xx^\top dP_n\right)^{-1} \int \frac{1}{v_{\hat{\theta}}(x)} xy dP_n \ ,$$

$$\beta(P_n) := \left(\int xx^\top dP_n\right)^{-1} \int xy dP_n \ ,$$

and

$$\beta_W(P_n) := \left(\int \frac{1}{w(x)} xx^\top dP_n\right)^{-1} \int \frac{1}{w(x)} xy dP_n \ .$$

Then,

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}_{\text{WLS}} - \beta_{\text{WLS}}(P_n)\right) - \sqrt{n}\left(\hat{\beta}_W - \beta_W(P_n)\right) &= (X_n^\top \hat{W}^{-1} X_n)^{-1} X_n^\top \hat{W}^{-1} \varepsilon_{\hat{W},n} \\ &\quad - (X_n^\top W^{-1} X_n)^{-1} X_n^\top W^{-1} \varepsilon_{W,n} \ .\end{aligned}$$

To show this quantity converges in probability to zero, it suffices to show that

$$\frac{1}{\sqrt{n}}\left(X_n^\top \hat{W}^{-1} \varepsilon_{\hat{W},n} - X_n^\top W^{-1} \varepsilon_{W,n}\right) \xrightarrow{P} 0$$

and

$$\frac{1}{n}\left(X_n^\top \hat{W}^{-1} X_n - X_n^\top W^{-1} X_n\right) \xrightarrow{P} 0 \ .$$

We can write the first expression as

$$\frac{1}{\sqrt{n}}\left[X_n^\top \left(\hat{W}^{-1} - W^{-1}\right)\varepsilon_{W,n} + X_n^\top \hat{W}^{-1} X_n \left(\beta_{\hat{W}}(P_n) - \beta_W(P_n)\right)\right] \ .$$

By the assumptions on sequences in C_P, $\sqrt{n}\left(\beta_{\hat{W}} - \beta_W\right) \xrightarrow{P} 0$. It will be seen later that $\frac{1}{n} X_n^\top \hat{W}^{-1} X_n \xrightarrow{P} \mathbb{E}(x^\top x / w(x))$, so the second tern in the above expression converges to zero in probability. The first term is

$$\frac{1}{\sqrt{n}} X_n^\top \left(\hat{W}^{-1} - W^{-1}\right)\varepsilon_{W,n} = \frac{1}{\sqrt{n}} \sum x_{n,i}^\top \left(\frac{1}{v_{\hat{\theta}}(x_{n,i})} - \frac{1}{v_{\theta_0}(x_{n,i})}\right)\varepsilon_{W,n,i}$$

which, as in Romano and Wolf (2017), can be written as $A + B$ where the j^{th} entry of A is

$$A_j = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{n,i,j}\varepsilon_{W,n,i} \sum_{l=1}^{K} r_{\theta_0,l}(x_{n,i})(\hat{\theta}_l - \theta_{0,l}) \ ,$$

and with probability tending to one,

$$|B_j| \leq \frac{1}{2\sqrt{n}} \left|\hat{\theta} - \theta_0\right|^2 \sum |x_{n,i,j}\varepsilon_{W,n,i} s_{\theta_0}(x_{n,i})| \ .$$

Because $n^{1/4}(\hat{\theta}_l - \theta_{0,l}) \xrightarrow{P} 0$, to show $A_j \xrightarrow{P} 0$, we only need to show that

$$n^{-3/4} \sum_{i=1}^{n} x_{n,i,j}\varepsilon_{W,n,i} r_{\theta_0,l}(x_{n,i}) \xrightarrow{P} 0$$

for each $l = 1, ..., K$. We will do this by showing that the mean and variance converge to zero. The variance converges to zero since

$$\text{var}_{P_n}\left(n^{-3/4} \sum_{i=1}^{n} x_{n,i,j}\varepsilon_{W,n,i} r_{\theta_0,l}(x_{n,i})\right) = n^{-1/2} \text{var}_{F_n}\left(x_{n,i,j}\varepsilon_{W,n,i} r_{\theta_0,l}(x_{n,i})\right)$$

and, by the assumptions on C_P, the sequence of variances $\text{var}_{P_n}\left(x_{n,i,j}\varepsilon_{W,n,i} r_{\theta_0,l}(x_{n,i})\right)$ is bounded. To show that the mean converges to zero, write

$$n^{-3/4} \sum_{i=1}^{n} x_{n,i,j}\varepsilon_{W,n,i} r_{\theta_0,l}(x_{n,i}) = n^{-3/4} \sum_{i=1}^{n} x_{n,i,j}\varepsilon_{n,i} r_{\theta_0,l}(x_{n,i}) + n^{-3/4} \sum_{i=1}^{n} (\varepsilon_{W,n,i} - \varepsilon_{n,i}) x_{n,i,j} r_{\theta_0,l}(x_{n,i}) \ .$$

The expectation of the first term converges to zero by assumption and the expectation of the second term converges to zero, since

$$\mathbb{E}_{P_n}\left(n^{-3/4} \sum_{i=1}^{n} x_{n,i,j}\varepsilon_{n,i} r_{\theta_0,l}(x_{n,i})\right) = \mathbb{E}_{P_n}\left(\frac{1}{n} X_{n,i} x_{n,i,j} r_{\theta_0,l}(x_{n,i})\right) n^{1/4}\left(\hat{\beta}(P_n) - \hat{\beta}_W(P_n)\right) \to 0 \ .$$

Similarly, since $\sqrt{n}\left|\hat{\theta}-\theta_0\right|^2 \xrightarrow{P} 0$, we have that $|B_j| \xrightarrow{P} 0$ provided $\frac{1}{n}\sum \left|x_{n,i,j}\varepsilon_{W,n,i}s_{\theta_0}(x_{n,i})\right| = O_p(1)$. As in the argument for A_j, this last sum has expectation tending to a constant, and variance tending to zero, and so it converges in probability to a constant.

Finally we must show that

$$\frac{1}{n}\left(X_n^\top \hat{W}^{-1}X_n - X_n^\top W^{-1}X_n\right) = \frac{1}{n}\sum x_i^\top x_{n,i}\left(\frac{1}{v_{\hat{\theta}}(x_{n,i})} - \frac{1}{v_{\theta_0}(x_{n,i})}\right)$$

converges in probability to zero. The argument proceeds as above.

Since $\sqrt{n}\big(\hat{\beta}_{\hat{W}} - \hat{\beta}_W\big)$ converges to zero in probability, but not necessarily almost surely, the empirical distribution functions $\hat{P}_n$ do not lie in C_P almost surely. However, it is easily seen that the empirical distribution functions satisfy the moment conditions on C_P in probability, so the asymptotic normality of the bootstrap distribution holds in probability. ■

Proof of Lemma 9.2 using the wild bootstrap. Let S' be the set on which $(S1)$–$(S4)$ hold as well as

(S5) $\frac{1}{n}\sum_{i=1}^n \left|x_i\hat{y}_i r_{\theta_{0,l}}(x)\right|^2 \to \mathbb{E}_P(|x_i y_i r_{\theta_0,l}(x)|^2)$ for each $i = 1, ..., p, l = 1, ..., d$,

(S6) $\frac{1}{n}\sum_{i=1}^n |x_i\hat{y}_i s_{\theta_0}(x)|^2 \to \mathbb{E}_P(|x_i y_i s_{\theta_0}(x)|^2)$ for each $i = 1, ..., p, l = 1, ..., d$, and

(S7) $n^{1/4}\big(\hat{\theta}^* - \theta_0\big)$ converges in probability to zero.

We will show that

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}^*_{\mathrm{WLS}} - \hat{\beta}_{\mathrm{WLS}}\right) - \sqrt{n}\left(\hat{\beta}^*_W - \hat{\beta}_W\right) =& \sqrt{n}\left[\left(X^\top W^{*-1}X\right)^{-1}X^\top W^{*-1}\varepsilon^* \right.\\ &\left. - \left(X^\top W^{-1}X\right)^{-1}X^\top W^{-1}\varepsilon^*\right] + \sqrt{n}\left(\hat{\beta}_{\mathrm{WLS}} - \hat{\beta}_W\right)\end{aligned}$$

converges to probability to zero, conditional on any sequence of $x's$ and $y's$ in S'.

By assumption, the second term converges to zero on S'. To show the first term converges in probability to zero, we will show that

$$\frac{1}{\sqrt{n}}\left(X_n^\top \hat{W}^{*-1}\varepsilon^* - X_n^\top W^{-1}\varepsilon^*\right) \xrightarrow{P} 0$$

and

$$\frac{1}{n}\left(X_n^\top \hat{W}^{*-1}X_n - X_n^\top W^{-1}X_n\right) \xrightarrow{P} 0 \ .$$

The first quantity can be written as

$$\frac{1}{\sqrt{n}}X_n^\top\left(\hat{W}^{-1} - W^{-1}\right)\varepsilon^* = \frac{1}{\sqrt{n}}\sum x_{n,i}^\top\left(\frac{1}{v_{\hat{\theta}^*}(x_{n,i})} - \frac{1}{v_{\theta_0}(x_{n,i})}\right)\varepsilon_i^*$$

which again can be written as $A + B$ where the j^{th} entry of A is

$$A_j := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{n,i,j} \varepsilon_i^* \sum_{l=1}^{K} r_{\theta_0,l}(x_{n,i})(\hat{\theta}_l^* - \theta_{0,l}) \ ,$$

and with probability tending to one,

$$|B_j| \leq \frac{1}{2\sqrt{n}} \left| \hat{\theta}^* - \theta_0 \right|^2 \sum |x_{n,i,j} \varepsilon_i^* s_{\theta_0}(x_{n,i})| \ .$$

By assumption $(S7)$, $n^{1/4}(\hat{\theta}_l^* - \theta_{0,l}) \xrightarrow{p} 0$. Further, for each l, $n^{-3/4} \sum_{i=1}^{n} x_{n,i,j} \varepsilon_i^* \sum_{l=1}^{K} r_{\theta_0,l}(x_{n,i})$ converges in probability to zero since it has mean zero and variance

$$\text{var}\left(n^{-3/4} \sum_{i=1}^{n} x_{n,i,j} \varepsilon_i^* r_{\theta_0,l}(x_{n,i}) \right) = n^{-3/2} \sum_{i=1}^{n} (x_{n,i,j} \hat{\varepsilon}_i r_{\theta_0,l}(x_{n,i}))^2$$

which converges to zero on S' by assumption $(S5)$. Consequently, A_j converges in probability to zero for each j. Similarly, B_j converges in probability to zero since $\sqrt{n}(\hat{\theta}_l^* - \theta_{0,l})^2$ converges in probability to zero, and $\frac{1}{n} \sum |x_{n,i,j} \varepsilon_i^* s_{\theta_0}(x_{n,i})|$ converges in probability to a constant.
The other convergence,

$$\frac{1}{n} \left(X_n^\top \hat{W}^{*-1} X_n - X_n^\top W^{-1} X_n \right) \xrightarrow{P} 0 \ ,$$

follows from a similar argument. ■

Proof of Lemma 3.1. We will first consider the estimate $\tilde{\theta}$ obtained by regressing $h_\delta(\varepsilon_i)$ on $g(x_i)$. By a similar argument to Lemma 9.1, $\sqrt{n}\left(\tilde{\theta}^* - \tilde{\theta}\right)$ is almost surely asymptotically normal. Consequently, $n^{1/4}\left(\tilde{\theta}^* - \tilde{\theta}\right)$ converges in conditional probability to zero, almost surely. We can express

$$\begin{aligned} n^{1/4}\left(\tilde{\theta} - \theta_0\right) &= n^{1/4}\left((G^\top G)^{-1} G^\top h - \theta_0\right) \\ &= n^{1/4}(G^\top G)^{-1} G^\top e \ . \end{aligned}$$

where G and h are the matrix and vector containing the $g(x_i)$ and $h_\delta(\varepsilon_i)$, respectively, and e is the vector with entries $e_i = h_\delta(y_i) - g(x)\theta_0$. Since $(\frac{1}{n} G^\top G)^{-1}$ converges almost surely to $\mathbb{E}(g(x_i)^\top g(x_i))$ and $n^{-3/4} G^\top e$ converges in almost surely to zero, $n^{1/4}\left(\tilde{\theta} - \theta_0\right)$ converges almost surely to zero.
Writing

$$n^{1/4}\left(\tilde{\theta}^* - \theta_0\right) = n^{1/4}\left(\tilde{\theta}^* - \tilde{\theta}\right) + n^{1/4}\left(\tilde{\theta} - \theta_0\right) \ ,$$

we see this quantity converges in conditional probability to zero, almost surely.
Now,

$$\hat{\theta}^* - \tilde{\theta}^* = \left(\frac{1}{n} \sum g(x_i^*) g^\top(x_i^*) \right)^{-1} \frac{1}{n} \sum g(x_i^*) \left(h_\delta(\hat{\varepsilon}_i^*) - h_\delta(\varepsilon_i^*) \right) \ .$$

It is easily seen that $\left(\frac{1}{n} \sum g(x_i^*) g^\top(x_i^*)\right)$ converges in conditional probability to $\mathbb{E}(g(x)g(x)')$ and $n^{-3/4} \sum g(x_i^*) \left(h_\delta(\hat{\varepsilon}_i^*) - h_\delta(\varepsilon_i^*) \right)$ converges in conditional probability to zero, almost surely. ■

Proof of Theorem 3.2. The bootstrap estimator $\hat{\Omega}^{*-1}_{1/w}\hat{\Omega}^{*}_{v/w^2}\hat{\Omega}^{*-1}_{1/w}$ converges in conditional probability to $\Omega^{-1}_{1/w}\Omega_{v/w^2}\Omega^{-1}_{1/w}$. As a consequence of Theorem 2, the bootstrap distribution of $\sqrt{n}R(\beta^*_{WLS} - \hat{\beta}_{\text{WLS}})$ approximates the distribution of $\sqrt{n}(R\hat{\beta} - q)$. It follows that the bootstrap distribution of W^*_n consistently approximates the distribution of W_n. Moreover, both the bootstrap distribution of M^*_n and the sampling distribution of M_n are asymptotically distributed as $\max_i |Z_i|$ where Z is a multivariate normal random variable with mean zero and covariance matrix $V\Omega^{-1}_{1/w}\Omega_{v/w^2}\Omega^{-1}_{1/w}V$, with V a diagonal matrix whose diagonal entries are equal to the square root of the diagonal entries of $\Omega^{-1}_{1/w}\Omega_{v/w^2}\Omega^{-1}_{1/w}$. The claims of the theorem now follow from Slutsky's Theorem. ■

Proof of Theorem 3.3 and Lemma 3.2. These claims follow from the same arguments as the wild bootstrap counterparts, but with $\hat{\varepsilon}_i$ replaced by ε_i. ■

Proof of Theorem 4.1. For almost all sequences $\{(x_i, y_i)\}$, $\widehat{\text{Avar}}(\hat{\beta}_{\text{OLS},k})^*$ converges to $\text{Avar}(\hat{\beta}_{\text{OLS},k})$ and $\widehat{\text{Avar}}(\hat{\beta}_{\text{WLS},k})$ converges to $\text{Avar}(\hat{\beta}_{\text{WLS},k})$ in conditional probability. The claim follows from applying Slutsky's theorem conditionally. ■

Proof of Theorem 4.2. Following the argument of Theorem 3.1 of Romano and Wolf (2017), we must only find the asymptotic joint distribution of $\sqrt{n}(\hat{\beta}_W - \beta)$ and $\sqrt{n}(\hat{\beta}_{\text{OLS}} - \beta)$ since $\sqrt{n}(\hat{\beta}_{\text{WLS}} - \hat{\beta}_W) \xrightarrow{P} 0$. We can write $\sqrt{n}(\hat{\beta}_W - \beta) = \left(\frac{1}{n}X^\top W^{-1}X\right)^{-1}\frac{1}{\sqrt{n}}X^\top W^{-1}\varepsilon$ and $\sqrt{n}(\hat{\beta}_{\text{OLS}} - \beta) = \left(\frac{1}{n}X^\top X\right)^{-1}\frac{1}{\sqrt{n}}X^\top\varepsilon$. Because

$$\left(\frac{1}{n}X^\top W^{-1}X\right)^{-1} \xrightarrow{P} \mathbb{E}\left(\frac{1}{w(x_i)}x_i^\top x_i\right)^{-1} = \Omega^{-1}_{1/w} \ ,$$

and

$$\left(\frac{1}{n}X^\top X\right)^{-1} \xrightarrow{P} \mathbb{E}\left(x_i^\top x_i\right)^{-1} = \Omega^{-1}_{1/1} \ ,$$

it is enough to find the joint limiting distribution of $\frac{1}{\sqrt{n}}X^\top W^{-1}\varepsilon$ and $\frac{1}{\sqrt{n}}X^\top\varepsilon$. These are scaled sums of i.i.d. mean zero random variables, so the Multivariate Central Limit Theorem gives

$$\sqrt{n}\begin{pmatrix} \frac{1}{n}X^\top W^{-1}\varepsilon \\ \frac{1}{n}X^\top\varepsilon \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbb{E}\left(x_i^\top x_i \frac{v(x_i)}{w^2(x_i)}\right) & \mathbb{E}\left(x_i^\top x_i \frac{v(x_i)}{w(x_i)}\right) \\ \mathbb{E}\left(x_i^\top x_i \frac{v(x_i)}{w(x_i)}\right) & \mathbb{E}\left(x_i^\top x_i v(x_i)\right) \end{pmatrix}\right) \ .$$

The claim follows from Slutsky's Theorem. ■

Proof of Theorem 4.3. An argument analogous to the proof of Theorem 3.1 to the one presented above shows that for any fixed λ, the bootstrap distribution of

$$\sqrt{n}(\lambda\hat{\beta}^*_{\text{WLS}} + (1-\lambda)\hat{\beta}^*_{\text{OLS}} - \lambda\hat{\beta}_{\text{WLS}} - (1-\lambda)\hat{\beta}_{\text{OLS}}) = \sqrt{n}(\hat{\beta}^*_\lambda - \hat{\beta}_\lambda) \ ,$$

is asymptotically normal with mean zero and covariance matrix $\text{Avar}(\hat{\beta}_\lambda)$ in probability.
It follows from the weak law of large numbers for triangular arrays that $\widehat{\text{Avar}}(\hat{\beta}_\lambda)^*$ converges in conditional probability to $\text{Avar}(\hat{\beta}_\lambda)$, almost surely. The second convergence follows from Slutsky's Theorem. ■

Proof of Theorem 4.4. We begin with the case where $\text{Avar}(\hat{\beta}_{\lambda,k})$ is non-constant. In order to show that $\sqrt{n}\big(\hat{\beta}_{\hat{\lambda}} - \beta\big) \xrightarrow{d} N\big(0, \text{Avar}(\hat{\beta}_{\lambda_0})\big)$, we will show that $\sqrt{n}\big(\hat{\beta}_{\hat{\lambda}_0} - \beta\big) - \sqrt{n}\big(\hat{\beta}_{\lambda_0} - \beta\big) \xrightarrow{P} 0$. Indeed,

$$\sqrt{n}\left(\hat{\beta}_{\hat{\lambda}_0} - \beta\right) - \sqrt{n}\left(\hat{\beta}_{\lambda_0} - \beta\right) = \sqrt{n}\left(\hat{\lambda}_0 - \lambda_0\right)\left[\hat{\beta}_{\text{OLS}} - \hat{\beta}_{\text{WLS}}\right]$$

which converges in probability to zero.

Theorem 4.3 gives that for any fixed λ, the bootstrap distribution of

$$\sqrt{n}(\lambda\hat{\beta}^*_{\text{WLS}} + (1-\lambda)\hat{\beta}^*_{\text{OLS}} - \lambda\hat{\beta}_{\text{WLS}} - (1-\lambda)\hat{\beta}_{\text{OLS}}) = \sqrt{n}(\hat{\beta}^*_\lambda - \hat{\beta}_\lambda)\ ,$$

is asymptotically normal with mean zero and covariance matrix $\text{Avar}(\hat{\beta}_\lambda)$ in conditional probability. To prove the convergence of the bootstrap distribution stated in the theorem, we will first show that the bootstrap distribution of $\sqrt{n}\left(\hat{\beta}^*_{\hat{\lambda}^*} - \hat{\beta}_{\hat{\lambda}^*}\right)$ is asymptotically normal with mean 0 and covariance matrix $\text{Avar}(\hat{\beta}_\lambda)$ in probability and then show that $\sqrt{n}\left(\hat{\beta}^*_{\hat{\lambda}^*} - \hat{\beta}_{\hat{\lambda}}\right) - \sqrt{n}\left(\hat{\beta}^*_{\hat{\lambda}^*} - \hat{\beta}_{\hat{\lambda}^*}\right) \xrightarrow{p} 0$ in probability.

To show the desired asymptotic normality of $\sqrt{n}\left(\hat{\beta}^*_{\hat{\lambda}^*} - \hat{\beta}_{\hat{\lambda}^*}\right)$, we will show

$$\sqrt{n}\left(\hat{\beta}^*_{\hat{\lambda}_0} - \hat{\beta}_{\lambda_0}\right) - \sqrt{n}\left(\hat{\beta}^*_{\hat{\lambda}^*} - \hat{\beta}_{\hat{\lambda}^*}\right) \xrightarrow{P} 0\ .$$

We can write

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}^*_{\hat{\lambda}_0} - \hat{\beta}_{\lambda_0}\right) - \sqrt{n}\left(\hat{\beta}^*_{\hat{\lambda}^*} - \hat{\beta}_{\hat{\lambda}^*}\right) =& \sqrt{n}(\hat{\lambda}^* - \lambda_0)\left[\hat{\beta}^*_{\text{WLS}} - \hat{\beta}_{\text{WLS}}\right] \\ &+ \sqrt{n}\left((1-\hat{\lambda}^*) - (1-\lambda_0)\right)\left[\hat{\beta}^*_{\text{OLS}} - \hat{\beta}_{\text{OLS}}\right]\ .\end{aligned}$$

Because $\sqrt{n}\big(\hat{\beta}^*_{WLS} - \hat{\beta}_{\text{WLS}}\big)$ and $\sqrt{n}\big(\hat{\beta}^*_{\text{OLS}} - \hat{\beta}_{\text{OLS}}\big)$ are asymptotically normal (in probability), the desired convergence follows from Slutsky's Theorem if we can show $\hat{\lambda}^* \xrightarrow{P} \lambda_0$. Note that $\hat{\lambda}^*$ is a continuous function of $\left[\hat{\Omega}^{*-1}_{1/w}\hat{\Omega}^*_{v/w^2}\hat{\Omega}^{*-1}_{1/w}\right]_{k,k}$, $\left[\hat{\Omega}^{*-1}_{1/w}\hat{\Omega}^*_{v/w}\hat{\Omega}^{*-1}_{1/1}\right]_{k,k}$, and $\left[\hat{\Omega}^{*-1}_{1/1}\hat{\Omega}^*_{v/1}\hat{\Omega}^{*-1}_{1/1}\right]_{k,k}$. Because these quantities converge in probability to the population versions almost surely, it follows from the continuous mapping theorem that $\hat{\lambda}^*$ converges in conditional probability to λ_0.

Similarly,

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}^*_{\hat{\lambda}^*} - \hat{\beta}_{\hat{\lambda}^*}\right) - \sqrt{n}\left(\hat{\beta}^*_{\hat{\lambda}^*} - \hat{\beta}_{\hat{\lambda}}\right) =& \sqrt{n}\left(\hat{\beta}_{\hat{\lambda}^*} - \hat{\beta}_{\hat{\lambda}_0}\right) \\ =& \sqrt{n}(\hat{\lambda}^* - \hat{\lambda}_0)\left[\hat{\beta}^*_{\text{WLS}} - \hat{\beta}_{\text{WLS}}\right] \\ \xrightarrow{P}& 0\end{aligned}$$

in conditional probability.

The case where $\text{Avar}(\hat{\beta}_{\lambda,k})$ is constant is similar, but follows from a simpler argument. ■

References

Angrist, J. D. and Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24:3–30.

Chesher, A. (1989). Hájek inequalities, measures of leverage, and the size of heteroskedasticity robust tests. *Econometrica*, 57:971–977.

Cribari-Neto, F. (2004). Asymptotic inference under heterskedasticty of unknown form. *Computational Statistics & Data Analysis*, 45:215–233.

Davidson, R. and Flachaire, E. (2008). The wild bootstrap, tamed at last. *Journal of Econometrics*, 146(1):162–169.

Flachaire, E. (2005). Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap. *Computational Statistics and Data Analysis*, 49(2):361–376.

Freedman, D. A. (1981). Bootstrapping regression models. *Annals of Statistics*, 9(6):1218–1228.

Godfrey, L. and Orne, C. (2004). Controlling the finite sample significance levels of heteroskedasticity-robust tests of several linear restrictions on regression coefficients. *Economics Letters*, 82:281–287.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

Janssen, A. (1999). Nonparametric symmetry tests for statistical functionals. *Mathematical Methods of Statistics*, 8:320–343.

Leamer, E. E. (2010). Tantalus on the road to asymptotia. *Journal of Economic Perspectives*, 24(2):31–46.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, New York, third edition.

Liu, R. Y. (1988). Bootstrap procedures under some non-i.i.d. models. *Annals of Statistics*, 16(4):1696–1708.

MacKinnon, J. G. (2012). Thirty years of heteroskedasticity-robust inference. In Chen, X. and Swanson, N. R., editors, *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, pages 437–461. Springer, New York.

MacKinnon, J. G. and White, H. L. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite-sample properties. *Journal of Econometrics*, 29:53–57.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Annals of Statistics*, 21(1):255–285.

Romano, J. P. and Wolf, M. (2017). Resurrecting weighted least squares. *Journal of Econometrics*, 197:1–19.

White, H. L. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test of heteroskedasticity. *Econometrica*, 48:817–838.

Wooldridge, J. M. (2012). *Introductory Econometrics*. South-Western, Mason, Ohio, fifth edition.

Wu, C. F. J. (1986). Jacknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics*, 14:1261–1350.